SMALL-SAMPLE PROPERTIES OF SEVERAL TWO-STAGE REGRESSION METHODS IN THE CONTEXT OF AUTO-CORRELATED ERRORS*

POTLURI RAO AND ZVI GRILICHES
University of Chicago

In a linear regression model, when errors are autocorrelated, several asymptotically efficient estimators of parameters have been suggested in the literature. In this paper we study their small sample efficiency using Monte Carlo methods.

While none of these estimators turns out to be distinctly superior to the others over the entire range of parameters, there is a definite gain in efficiency to be had from using some two-stage procedure in the presence of moderate high levels of serial correlation in the residuals and very little loss from using such methods when the true ρ is small. Where computational costs are a consideration a mixed strategy of switching to a second stage only if the estimated $\hat{\rho}$ is higher than some critical value is suggested and is shown to perform quite well over the whole parameter range.

1. INTRODUCTION

IN THE standard linear regression model autocorrelation of the disturbances leads to inefficient but still unbiased estimates of the coefficients. Since the autocorrelation parameters of the disturbances are rarely known *a priori*, one cannot use minimum variance Generalized Least Squares methods directly. Several two-stage estimation procedures have been suggested in the literature and it has also been shown that if the first-stage estimates of the variance-covariance matrix of the errors are consistent, the resulting second stage estimators are also consistent and *asymptotically* efficient. Little is known, however, about the small sample properties of such estimators. Does the use of a consistent but often relatively poor (high variance) estimator of the serial correlation coefficient from the computed first stage residuals "really" reduce the variance of the second stage estimators? The purpose of this paper is to report on a Monte-Carlo study of this question. Before presenting the results of the investigation we first describe the model, its theoretical properties, and the estimates compared in this study.

2. THEORETICAL CONSIDERATIONS

We consider the following model:

$$y_t = \beta x_t + u_t,$$
$$x_t = \lambda x_{t-1} + v_t,$$
$$u_t = \rho u_{t-1} + w_t,$$
$$E(v_t) = E(w_t) = E(v_t w_t) = E(w_t w_{t-1}) = E(v_t v_{t-1}) = 0,$$
$$E(v_t^2) = \sigma_v^2, \qquad E(w_t^2) = \sigma_w^2, \qquad |\lambda| < 1, \qquad |\rho| < 1$$

* This work has been supported by grants from the Ford and National Science Foundations.

and

$$t = 1, 2, 3, \cdots, T.$$

Assume that the stationary processes generating u and x started in the past and continue to operate. This is similar to assuming that the initial values of u and x are drawn from a normal population with zero means and variances $\sigma_v^2 = \sigma_w^2/(1-\rho^2)$ and $\sigma_x^2 = \sigma_v^2/(1-\lambda^2)$ respectively. We consider all the variables as deviations from their means and hence do not include a constant term explicitly in the model.

The covariance matrix of the error vector is:

$$E(uu') = R = \sigma_u^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdot & \cdot & \cdot & \rho^{T-1} \\ \rho & 1 & \rho & \cdot & \cdot & \cdot & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdot & \cdot & \cdot & \rho^{T-3} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdot & \cdot & \rho & 1 \end{bmatrix}$$

There are several ways of estimating β from a given sample:

1) *Generalized Least Squares (G.L.S.)*:

This procedure is possible only when ρ is known. The G.L.S. estimator is given by

$$b_G = (x'R^{-1}x)^{-1}x'R^{-1}y,$$

and

$$V(b_G) = (x'R^{-1}x)^{-1}.$$

It is a minimum variance linear unbiased estimator. A large sample approximation to its variance is obtained as follows:

$$x'R^{-1}x = (x_1 x_2 \cdots x_T) \cdot \frac{1}{(1-\rho^2)\sigma_u^2} \cdot \begin{bmatrix} 1 & -\rho & 0 & \cdot & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdot & 0 \\ 0 & -\rho & 1+\rho^2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & -\rho & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_T \end{bmatrix}$$

$$= \frac{1}{(1-\rho^2)\sigma_u^2} \cdot \left(\sum_1^T x_t^2 + \rho^2 \sum_2^{T-1} x_t^2 - 2\rho \sum_2^T x_t x_{t-1} \right)$$

$$= \frac{1}{(1-\rho^2)\sigma_u^2} \cdot \left(\sum_1^T x_t^2 + \rho^2 \sum_2^{T-1} x_t^2 - 2\rho\lambda \sum_1^{T-1} x_t^2 - 2\rho \sum_2^T v_t x_{t-1} \right)$$

For large samples we may ignore the last term of the above expression as x_t and v_{t+1} are not correlated. Thus

$$x'R^{-1}x \simeq \frac{1}{(1-\rho^2)\sigma_u^2} \cdot \left(1 + \rho^2 - 2\rho\lambda - \frac{\rho^2 x_T^2 + \rho^2 x_1^2 - 2\rho\lambda x_T^2}{\sum_1^T x_t^2} \right) \sum_1^T x_t^2.$$

For reasonably large T the last term in the parenthesis approaches zero and as an approximation:[1]

$$(x'R^{-1}x)^{-1} \simeq \frac{\sigma_u^2}{\sum_1^T x_t^2} \cdot \frac{1-\rho^2}{1+\rho^2-2\rho\lambda}$$

2) *Ordinary Least Squares (O.L.S.)*:

$$b_0 = (x'x)^{-1}x'y$$

The estimator b_0 is unbiased and its variance is:

$$V(b_0) = E(\Sigma x_t u_t / \Sigma x_t^2)^2$$
$$= \left(1 \Big/ \sum_1^T x_t^2\right)^2 \cdot E\left(\sum_1^T x_t u_t\right)^2$$

For large samples (see equation 1 of the appendix) we can show that the following approximation holds:[2]

$$V(b_0) \simeq \frac{\sigma_u^2}{\Sigma x^2} \cdot \frac{1+\rho\lambda}{1-\rho\lambda}$$

Hence the efficiency of O.L.S. for reasonably large samples is approximately[3]

$$\text{Eff} = V(b_G)/V(b_0) \simeq \frac{1-\rho\lambda}{1+\rho\lambda} \cdot \frac{1-\rho^2}{1+\rho^2-2\rho\lambda}.$$

Since ρ is in general unknown, several methods of estimating and using the result to improve on the efficiency of O.L.S. have been suggested in the literature. We shall discuss only four of these: three two-stage methods (Cochrane-Orcutt, Durbin, and Prais-Winsten) and one non-linear approach.

3) *The Cochrane and Orcutt estimator (C.O.)*:

In their 1949 paper Cochrane and Orcutt [2] suggested the following estimator for models with autocorrelated errors:[4] Let ρ_0 be a consistent estimate of ρ from the residuals of O.L.S. Define,

$$z_t = y_t - \rho_0 y_{t-1}$$

and

$$q_t = x_t - \rho_0 x_{t-1}$$

---

[1] We are ignoring here a term that is approximately equal to $-2\rho(\rho-\lambda)/T$. It is far from negligible for small T and ρ and λ of opposite sign.

[2] This formula again overestimates the variance for low T and high ρ and λ.

[3] This expression is the same as the one derived by Johnston [8], p. 191, only when the independent variable is either serially uncorrelated ($\lambda=0$) or has the same autocorrelation coefficient as the errors ($\lambda=\rho$). A plot of this expression for selected values of λ is presented in Malinvaud [10], p. 439.

[4] Actually, Cochrane and Orcutt do not recommend the use of this estimator because of the downward bias in ρ_0. They also suggest the possibility of iterating several times more. Nevertheless, since they seem to be the first to mention such an estimator, we associate their names with it.

then

$$z_t = \beta q_t + w_t,$$

and

$$b_c = (q'q)^{-1}q'z \qquad t = 2, 3, \cdots, T.$$

When $\rho_0 = \rho$, the w's are serially independent and the O.L.S. estimator of β based on the above equation has minimum variance.[5] But ρ_0 is estimated from the residuals and will therefore usually not equal ρ. Whenever ρ_0 deviates from ρ, the C.O. estimator is not the minimum variance estimator. Lower bounds on the efficiency of such an estimator have been investigated by Watson and Hannan [15]. It has not been established, however, whether this estimator is better than O.L.S., on the average in *small samples.*

The conventional estimator of ρ from the residuals of O.L.S. is

$$\hat{\rho} = \sum_2^T e_t e_{t-1} \Big/ \sum_2^T e_t^2,$$

where e_t is the calculated O.L.S. residual for period t.[6]

The first term in the Taylor expansion of $E(\hat{\rho})$ gives terms up to the order of $(1/T)$.

$$E(\hat{\rho}) \simeq E(\Sigma e_t e_{t-1})/E(\Sigma e_t^2).$$

Using equations 2 and 3 of the appendix we can show that in our model

$$E(\hat{\rho}) \simeq \rho - \frac{\rho + \lambda}{T - 1 - \dfrac{1 + \rho\lambda}{1 - \rho\lambda}}.$$

The bias of $\hat{\rho}$ is a function of ρ, T and λ. It is often stated that $\hat{\rho}$ is biased towards zero. This is strictly true only if the autocorrelation of the independent variable is of the same sign as that of the disturbances.[7]

4) *The Durbin estimator*:

Durbin [4] suggested an estimator which is essentially the same as the C.O. estimator except that ρ_0 is obtained as the coefficient of y_{t-1} in the following regression equation.

$$y_t = \rho y_{t-1} + \beta x_t - \beta\rho x_{t-1} + w_t \qquad t = 2, 3, \cdots, T.$$

---

[5] Except for "end-effects" to be discussed below.

[6] So defined, $\hat{\rho}$ can occasionally exceed unity. When this happens several of the estimators we discuss are not defined. In actual computations we have, therefore, set $\hat{\rho}$ equal to $+1$ or -1, depending on the computed value, whenever the latter exceeded unity.

[7] For the case of k orthogonal independent variables, all generated by first order Markov schemes, we have similarly

$$E(\hat{\rho}) \simeq \rho - \frac{k\rho + \sum_1^k \lambda_i}{T - 1 - \sum_1^k \dfrac{1 + \rho\lambda_i}{1 - \rho\lambda_i}}$$

In this formulation the w's are serially independent and the O.L.S. estimate of the coefficient of y_{t-1} is consistent. Durbin also proves that the resulting second stage estimator of β is consistent and asymptotically efficient.

5) *The Prais-Winsten estimator (P.W.)*:

Neither the C.O. or the Durbin procedure is strictly speaking equivalent to the G.L.S. procedure even when $\hat{\rho}=\rho$. Both are based on a transformation that reduces the sample size from T to $T-1$. While this does not matter in large samples, it may make a significant difference in small samples. In an unpublished Cowles Foundation Discussion Paper, Prais and Winsten [12] pointed out in 1954 that the correct diagonalizing transformation matrix is not of the $T-1$ by T form

$$\begin{bmatrix} -\rho & 1 & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & -\rho & 1 & 0 & \cdot & \cdot & \cdot \\ 0 & 0 & -\rho & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

implied in the C.O. and Durbin procedures, but rather the T by T matrix

$$\begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdot & \cdot & \cdot \\ -\rho & 1 & 0 & \cdot & \cdot & \cdot \\ 0 & -\rho & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

with the first observation getting the weight of $\sqrt{1-\rho^2}$ instead of being "thrown away." They also pointed out that the loss in efficiency entailed in the usual procedure depends critically on how different the beginning x value is from the average, and that it could be quite high for trend like x's. In practice, the P.W. estimator is equivalent to the C.O. estimator except for the use of one additional observation, the first, with the weight $\sqrt{1-\rho^2}$.[8]

6) *A non-linear estimator*:

In this estimation procedure both β and ρ are estimated simultaneously using the following equation

$$y_t = \rho y_{t-1} + \beta x_t - \beta\rho x_{t-1} + w_t \qquad t = 2, 3, \cdots, T$$

but imposing the non-linear constraint

$$\widehat{\beta\rho} = \hat{\beta}\cdot\hat{\rho}$$

on the estimates.

Provided that the estimation procedure finds the absolute minimum of the residual sum of squares with respect to β and ρ, the resulting estimates have been shown to be maximum likelihood and hence asymptotically efficient by

[8] This was also recognized, implicitly, by Cochrane and Orcutt [2] in their original paper. See the discussion in their Appendix.

Hildreth [6] and Dhrymes [3]. Since it is not assured that the sample likelihood function has only one local maximum, Hildreth and Lu [7] and Dhrymes suggest computational procedures based on scanning over the whole range of ρ from -1 to $+1$ to assure the finding of the global maximum of the likelihood function. Their procedures, however, are not well adapted to Monte Carlo experiments. On the other hand, "standard" Newton-Raphson and Modified Gauss-Newton procedures[9] tried by us first failed to converge in many experiments. An investigation into the numerical properties of these procedures revealed that this failure was being caused by the size of the "step" taken by these procedures at each iteration. An alternative procedure in which an optimal "step" size is computed anew at each iteration on the basis of a parabolic approximation was developed by one of the authors.[10] It increased the speed of convergence greatly and converged almost always.

3. DESIGN OF THE EXPERIMENT

The following model was used to generate observations for the sampling experiment:

$$y_t = 0 + 1.0x_t + u_t,$$
$$x_t = \lambda x_{t-1} + v_t,$$
$$u_t = \rho u_{t-1} + w_t, \qquad t = 1, 2, \cdots, 20.$$

For any given ρ and λ we drew 50 samples with independent x and u series.[11] The initial values for x and u were drawn from normal populations with zero means and corresponding variances. v and w were drawn independently from normal populations with zero means and variances σ_v^2 and σ_w^2 respectively. The variances of v and w were adjusted so as to make the true square of the correlation coefficient between x and y equal to 0.9. Since we are interested only in the relative efficiency of the various estimators we set the coefficient of x to unity.

The various two-stage and non-linear estimation procedures were compared at intervals of 0.1 for ρ and 0.2 for λ. We restricted our study to stationary series only, and hence limited ρ and λ to the range $-1<\rho, \lambda<1$.

4. SMALL SAMPLE PROPERTIES OF $\hat{\rho}$

Since the performance of the various two-stage estimators depends heavily on the quality of $\hat{\rho}$ used in the second stage, it will help our understanding of what follows to review briefly the small sample performance of the three estimators of ρ used in this study. Both the C.O. and P.W. methods use as their $\hat{\rho}$ the serial correlation coefficient of the O.L.S. residuals, the Durbin procedure

[9] See Hartley [5].

[10] This procedure is described and compared to the Gauss-Newton alternatives in Rao [13]. It proved efficient even in many parameter models with the kind of non-linearity in the parameters discussed in this paper. Of course, there are many elaborate non-linear procedures designed to minimize complex functions and based on expensive algorithms. The simple procedure alluded to above appears, however, to be quite adequate and efficient for problems of non-linearity commonly faced in serial correlation and distributed lag models. An IBM 7094 computer program is available from the CMSBE, Graduate School of Business, University of Chicago.

[11] The results reported below are not sensitive to the particular procedure used in choosing the x's. In an earlier draft of this paper we kept the x series fixed for a given λ, drawing only a new u series for each sample. We did not notice any significant differences in the results of the two versions.

TABLE 1. REGRESSIONS OF MEAN BIAS OF DIFFERENT $\hat{\rho}$'S ON ρ AND λ

Estimator	Coefficients of			Standard error of regression
	Constant	ρ	λ	
		Non-positive values of ρ		
O.L.S.	−.048	−.195	−.055	.022
	(.004)	(.008)	(.004)	
Durbin	.001	−.106	−.052	.024
	(.004)	(.008)	(.004)	
Non-Linear	−.053	−.154	−.012	.039
	(.007)	(.014)	(.007)	
		Positive values of ρ		
O.L.S.	−.036	−.247	−.057	.029
	(.006)	(.010)	(.005)	
Durbin	−.018	−.057	−.042	.028
	(.006)	(.010)	(.005)	
Non-Linear	−.040	−.188	−.030	.031
	(.007)	(.011)	(.005)	

Dependent variable: $\hat{\rho}-\rho$. Numbers in parentheses are the computed standard errors of the respective coefficients. The range of the sample is −.9 to .99 for ρ and −.8 to .99 for λ.

uses the coefficient of y_{t-1} in the expanded equation as its estimate of ρ; and the non-linear procedure estimates ρ and β simultaneously. All these estimators of ρ are biased in small samples. No exact analytical expressions are available to compare their relative biases (a large sample expression for the bias of the O.L.S. residuals based $\hat{\rho}$ is given in this paper) and they have thus to be inferred from the Monte-Carlo results.

The Monte Carlo information on the bias of the three estimators of ρ is summarized in a set of regressions presented in Table 1. The dependent variable in each of these regressions is the average bias (averaged over 50 samples) at each set of ρ and λ. There are 200 such sets in total, but because the results appeared to be sensitive to the sign of ρ, these regressions were run separately for non-positive values of ρ and for $\rho>0$. Each regression is thus based on 100 observations, and each observation is an average of 50 independent samples for a particular ρ and λ.[12]

Table 1 indicates that none of the estimators has a uniformly lower bias. The Durbin estimator is significantly less biased than the other two for positive values of ρ. For small negative values of ρ the O.L.S. residuals based $\hat{\rho}$ has a somewhat smaller bias while the non-linear estimator has the lowest bias for large negative values of ρ. But the superiority of the latter two estimators over the Durbin estimator in the range of negative ρ's is only slight, while the Durbin estimator is significantly less biased for positive ρ.

[12] These regressions summarize the individual results quite well, the R^2's in the $\hat{\rho}$ dependent form being all in the .98–.99 range.

TABLE 2. RELATIVE MEAN SQUARE ERRORS OF DIFFERENT ESTIMATORS OF ρ: AVERAGES OVER ALL λ

ρ	MSE $\hat{\rho}$ O.L.S. / MSE $\hat{\rho}$ Durbin	MSE $\hat{\rho}$ O.L.S. / MSE $\hat{\rho}$ Non-linear	MSE $\hat{\rho}$ Durbin / MSE $\hat{\rho}$ Non-linear
0.9	4.53	1.20	0.28
0.8	2.52	1.17	0.49
0.7	1.88	1.15	0.63
0.6	1.54	0.98	0.65
0.5	1.45	1.06	0.74
0.4	1.29	0.93	0.75
0.3	1.07	0.88	0.83
0.2	0.99	0.88	0.86
0.1	0.99	0.93	0.82
0.0	0.86	0.88	0.92
−0.1	0.94	0.84	0.90
−0.2	0.85	0.85	1.00
−0.3	0.84	0.90	1.08
−0.4	0.86	0.93	1.10
−0.5	0.86	0.98	1.14
−0.6	0.86	1.03	1.20
−0.7	0.93	1.11	1.19
−0.8	1.04	1.43	1.39
−0.9	1.05	1.29	1.22

The performance of a particular $\hat{\rho}$ depends, however, not only on its average bias but also on its variance. Often a less biased estimator may have a higher variance cancelling much of the gain from the reduction in bias. To investigate this we computed the mean square error (M.S.E.) for each $\hat{\rho}$ over the 50 samples at each ρ and λ. To convert them into comparable units we divided them by the M.S.E. for the other $\hat{\rho}$'s, and, since such ratios did not appear to be very sensitive to λ, we averaged them over all (10) λ values. These average M.S.E. ratios are reported in Table 2. They again indicate that the Durbin $\hat{\rho}$ is significantly better for high positive ρ, while at the same time not being distinctly inferior to the other two methods for negative ρ's. Again, the non-linear estimator is better only for large negative ρ's. It should come then as no surprise if the Durbin second stage estimator of β does quite well in the comparisons to follow.

5. COMPARISON OF THE VARIOUS ESTIMATORS OF β

Table 3 summarizes the major results of this study. It presents the average performances (averaged over all λ) of all the estimators relative to the unattainable GLS estimator, based on a knowledge of the true ρ's, and shows that no estimator has a uniformly lower M.S.E. for all values of ρ and λ. We know that the O.L.S. estimator is not efficient in large samples for $\rho \neq 0$. We find that in small samples ($T=20$) the O.L.S. estimator is less efficient than all the other methods considered for moderate and high values of $\rho(|\rho| \geq .3)$. All the other estimators are not very far apart in their performance. The non-linear estimator is somewhat inferior to the two-stage estimators, while the C.O. estimator is

TABLE 3. AVERAGE RELATIVE EFFICIENCY OF DIFFERENT ESTIMATORS OF β FOR ALL VALUES OF λ

AVERAGE RATIOS OF MEAN SQUARE ERRORS

RHO	GLS OLS	GLS PW	GLS CO	GLS DURBIN	GLS NONLINEAR
0.99	0.15	0.65	0.77	0.95	0.79
0.9	0.31	0.79	0.80	0.95	0.76
0.8	0.35	0.81	0.78	0.90	0.77
0.7	0.47	0.86	0.80	0.96	0.83
0.6	0.56	0.87	0.83	0.89	0.78
0.5	0.74	0.90	0.86	0.90	0.85
0.4	0.79	0.88	0.83	0.84	0.76
0.3	0.83	0.93	0.84	0.86	0.79
0.2	0.94	0.95	0.92	0.95	0.90
0.1	0.99	0.93	0.86	0.87	0.82
0.0	1.00	0.94	0.90	0.90	0.84
-0.1	0.97	0.95	0.89	0.89	0.83
-0.2	0.96	0.92	0.83	0.84	0.79
-0.3	0.90	0.94	0.92	0.91	0.78
-0.4	0.80	0.97	0.90	0.88	0.81
-0.5	0.67	0.91	0.85	0.84	0.80
-0.6	0.61	0.92	0.82	0.83	0.79
-0.7	0.43	0.94	0.92	0.93	0.80
-0.8	0.31	0.86	0.83	0.88	0.79
-0.9	0.18	0.89	0.87	0.88	0.85

slightly inferior to the other two-stage estimators. A more detailed pairwise comparison of the various estimators follows below.

Table 4 compares in greater detail the performance of the O.L.S. estimator to one of the better two stage methods, the Prais-Winsten one. O.L.S. is distinctly inferior to the P.W. estimator for high (absolute value) ρ's, though it is not too bad when λ is high (and hence the x's are "smooth").[13] For small values of $\rho(|\rho| < .2)$ the O.L.S. estimator is actually more efficient than the P.W. even though the latter is based on a two stage procedure. To understand this point let us treat the O.L.S. estimator as a particular case of the P.W. with the $\hat{\rho}$ coming from a population with zero mean and zero variance. Whenever the estimated $\hat{\rho}$ deviates from its true value the P.W. estimator loses efficiency, and the loss in efficiency is a function of $\hat{\rho}$, ρ, and λ. In the case of the

[13] This is consistent with the argument presented by Chipman [1] for the efficiency of O.L.S. in the case of "smooth" x's.

O.L.S. estimator all the sample estimates are less efficient for non-zero ρ. But for the P.W. estimator the loss in efficiency has a distribution, because $\hat{\rho}$ comes from a population with non-zero variance. Table 4 shows that for small values of ρ the mean loss in efficiency is larger for the P.W. than for the O.L.S. estimator.

For small values of ρ efficiency of the O.L.S. estimator of β is quite high and increases with the autocorrelation of the independent variable. Of course, it is not obvious what is a "high" or a "low" efficiency without reference to some kind of loss function. Perhaps the following argument has some intuitive appeal:

A relative efficiency of 0.7 implies that one could reduce the standard error of β by about 16 per cent on the average by switching to the more efficient estimation method. This would raise a t-ratio from say 1.5 to 1.7. Such an improvement appears to us to be just on the "margin" for many econometric problems. One would like to gain "more" from a more efficient procedure to

TABLE 4. RELATIVE EFFICIENCY OF THE O.L.S. COMPARED TO THE P.W. ESTIMATOR OF β [M.S.E.(P.W.)/M.S.E.(O.L.S.)]

	LAMBDA									
RHO	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	0.99
0.99	0.19	0.32	0.14	0.29	0.14	0.22	0.21	0.25	0.23	0.41
0.9	0.80	0.51	0.23	0.27	0.30	0.23	0.29	0.30	0.29	0.54
0.8	0.47	0.60	0.25	0.38	0.36	0.44	0.43	0.45	0.55	0.37
0.7	1.01	0.39	0.52	0.40	0.36	0.56	0.60	0.54	0.53	0.46
0.6	0.63	0.81	0.57	0.43	0.70	0.46	0.77	0.74	0.79	0.54
0.5	1.08	1.15	0.72	0.77	0.67	0.76	0.81	0.90	0.70	0.60
0.4	0.86	0.95	0.73	0.91	0.79	0.97	0.91	1.02	0.97	0.81
0.3	1.08	0.75	1.01	0.78	0.79	1.01	0.98	0.84	0.83	0.89
0.2	0.97	1.01	0.92	0.83	1.07	0.92	0.97	1.04	1.11	0.97
0.1	1.16	1.12	0.99	0.96	1.05	1.08	1.12	1.09	1.03	1.03
0.0	1.08	1.24	1.17	1.08	0.87	0.95	1.07	1.00	1.10	1.22
-0.1	1.01	1.13	1.13	1.05	1.03	1.08	0.97	1.06	0.96	0.87
-0.2	1.02	1.07	1.12	1.01	1.31	1.18	0.85	0.97	0.97	0.96
-0.3	1.02	0.88	0.95	0.93	0.98	0.97	0.90	0.99	1.00	0.94
-0.4	0.78	1.08	1.00	0.67	0.59	0.71	0.61	0.95	1.01	0.83
-0.5	0.86	0.90	0.68	0.70	0.57	0.71	0.74	0.61	0.68	0.90
-0.6	0.71	0.69	0.64	0.98	0.65	0.40	0.55	0.66	0.82	0.48
-0.7	0.58	0.49	0.38	0.43	0.49	0.41	0.51	0.45	0.39	0.46
-0.8	0.41	0.30	0.33	0.38	0.19	0.27	0.55	0.50	0.27	0.41
-0.9	0.23	0.12	0.11	0.16	0.10	0.14	0.21	0.34	0.26	0.31

TABLE 5. RELATIVE EFFICIENCY OF THE P.W. ESTIMATOR COMPARED TO THE C.O. ESTIMATOR OF β [M.S.E.(C.O.)/M.S.E.(P.W.)]

	LAMBDA									
RHO	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	0.99
0.99	0.90	0.76	0.90	0.87	0.83	0.72	0.88	0.87	0.84	0.89
0.9	1.01	1.07	1.02	0.97	0.95	0.96	1.00	0.95	0.92	0.93
0.8	1.01	1.02	0.98	0.94	1.05	1.13	0.98	1.00	1.00	1.39
0.7	0.99	1.13	1.05	1.06	1.00	1.01	0.94	1.11	1.08	1.44
0.6	1.01	1.08	0.89	0.95	1.04	0.98	0.99	1.15	1.16	1.47
0.5	1.05	1.18	0.99	0.98	0.97	1.10	1.05	0.97	1.00	1.22
0.4	1.08	1.05	1.00	0.99	1.07	1.15	1.01	1.16	1.08	1.11
0.3	1.02	1.25	1.05	1.07	1.04	1.09	1.02	1.23	1.18	1.21
0.2	1.04	1.11	0.96	1.00	1.06	0.94	0.96	1.10	1.16	1.10
0.1	1.14	1.08	1.07	0.90	1.12	1.07	1.07	1.06	1.08	1.35
0.0	1.10	1.32	1.02	1.00	0.99	1.19	1.04	0.98	1.01	0.89
-0.1	1.16	1.10	1.00	1.05	1.08	1.15	1.04	1.11	1.06	0.98
-0.2	1.06	1.26	1.03	1.05	1.09	1.04	1.02	1.14	1.17	1.29
-0.3	1.05	1.09	0.95	1.01	1.04	0.96	0.96	1.01	1.11	1.12
-0.4	1.24	1.04	1.01	1.02	1.11	1.04	0.97	0.99	1.11	1.43
-0.5	1.10	0.98	1.05	1.04	1.08	1.02	1.06	0.97	1.06	1.39
-0.6	1.19	1.02	1.02	1.08	1.00	1.04	1.08	1.10	1.13	1.63
-0.7	1.07	0.92	0.99	1.02	0.94	0.97	0.98	1.00	0.95	1.46
-0.8	1.08	1.03	0.96	0.99	1.07	1.08	0.99	0.92	1.03	1.39
-0.9	1.01	0.88	0.95	1.00	1.02	0.97	1.02	1.00	1.04	1.41

justify the more complex computation and interpretation. If one accepts this rule of thumb, one would not switch from the O.L.S. to the P.W. estimator in small samples unless one had good reasons to believe that true ρ is equal to 0.4 or higher.

Both the P.W. and the C.O. are two-stage procedures using $\hat{\rho}$ from the O.L.S. except that the P.W. uses an extra observation at the second stage. Does the additional observation increase the efficiency of the P.W.? To answer this question we computed the relative efficiency of the P.W. compared to the C.O. in Table 5. The P.W. is more efficient than the C.O. except for values of ρ close to unity. When there is a strong trend in the errors the P.W. looses efficiency by including the first observation instead of "throwing it away" as in the C.O.

In interpreting the results of Table 5 we should not forget that the P.W. has an extra degree of freedom over the C.O. When we adjust for this additional

degree of freedom the gain in efficiency disappears except for values of λ close to unity. That is, when the independent variable is strongly trending information contained in the first observation does increase the efficiency of the second stage of the P.W. estimator above and beyond what one would get just by adding an additional observation to the sample.

The Durbin estimator and the C.O. are the same procedures except that one uses for $\hat{\rho}$ the coefficient of y_{t-1} in the expanded equation while the other is based on O.L.S. residuals. We have already noticed that neither of the ρ estimators has smaller M.S.E. over the entire parameter space. For a given parameter set we expect the least M.S.E. $\hat{\rho}$ to give the best results. To verify this point we computed the relative efficiency of the Durbin estimator compared to the C.O. in Table 6. It shows that the Durbin estimator is more efficient than the C.O. for ρ larger than 0.3 and is consistent with the findings of Table 2 which indicated that for ρ larger than 0.3 the Durbin $\hat{\rho}$ has the smaller M.S.E.

TABLE 6. RELATIVE EFFICIENCY OF THE DURBIN ESTIMATOR COMPARED TO THE C.O. ESTIMATOR OF β
(M.S.E.(C.O.)/M.S.E.(Durbin))

	LAMBDA									
RHO	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	0.99
0.99	1.04	1.03	1.02	1.42	1.06	1.13	1.28	1.83	1.17	1.84
0.9	0.97	0.91	1.03	1.60	1.15	0.95	1.19	1.50	1.65	1.53
0.8	1.03	1.01	1.17	1.13	1.32	1.26	1.17	1.23	1.00	1.32
0.7	1.00	1.06	1.10	1.00	1.11	1.07	1.09	1.06	1.22	1.24
0.6	1.04	1.03	0.98	1.04	1.09	1.09	1.21	1.00	1.05	1.29
0.5	0.97	1.04	1.07	1.13	1.03	1.00	1.03	0.91	1.18	1.28
0.4	1.04	0.99	1.14	0.96	1.08	1.02	0.94	0.96	0.99	1.03
0.3	1.00	1.05	1.06	1.04	1.05	0.96	0.99	0.99	1.08	1.04
0.2	1.03	1.04	0.99	1.02	1.01	1.06	1.08	1.01	0.94	1.10
0.1	1.02	1.02	1.00	1.00	0.99	1.02	1.01	1.05	0.95	1.02
0.0	1.00	1.00	1.03	0.94	1.01	0.98	0.99	0.94	0.97	1.10
-0.1	0.96	1.06	1.01	0.94	1.05	0.98	0.97	0.97	0.97	1.07
-0.2	1.10	1.01	1.04	1.03	0.97	0.99	0.99	0.93	1.03	0.94
-0.3	0.92	0.99	0.99	0.98	0.97	1.08	1.00	1.00	0.92	1.02
-0.4	0.99	1.02	0.95	1.03	0.98	0.95	0.96	0.93	1.00	0.97
-0.5	1.00	1.03	0.93	0.98	0.97	1.04	0.95	1.00	0.99	1.01
-0.6	0.94	1.12	1.07	0.95	1.03	1.07	1.00	0.98	0.99	0.98
-0.7	1.05	1.04	0.98	0.98	1.03	0.93	1.01	1.04	1.03	0.98
-0.8	1.18	1.09	1.34	1.04	1.06	1.03	1.02	0.98	1.03	1.00
-0.9	1.32	1.24	0.99	0.93	0.98	0.99	0.98	1.00	0.97	1.02

TABLE 7. RELATIVE EFFICIENCY OF THE P.W. ESTIMATOR WITH THE DURBIN $\hat{\rho}$ COMPARED TO THE DURBIN ESTIMATOR

Efficiency = M.S.E.(Durbin)/M.S.E.(P.W. with Durbin $\hat{\rho}$)

ρ	λ								
	−0.8	−0.6	−0.4	−0.2	0.0	0.2	0.4	0.6	0.8
0.9	1.07	1.09	1.03	0.97	0.98	0.99	1.01	0.90	1.08
0.8	1.02	1.01	1.02	0.96	1.11	1.15	0.95	1.02	1.04
0.7	0.99	1.17	1.06	1.10	1.01	1.01	0.98	1.14	1.08
0.6	1.01	1.07	0.89	0.96	1.05	0.96	0.99	1.18	1.17
0.5	1.08	1.17	0.99	0.99	0.96	1.13	1.05	0.98	1.01
0.4	1.08	1.05	1.00	1.00	1.13	1.17	1.01	1.18	1.09
0.3	1.02	1.27	1.05	1.08	1.05	1.10	1.03	1.25	1.18
0.2	n.c.	1.11	0.96	1.00	1.06	0.95	0.96	1.11	1.18

n.c. =not computed.

We noted above that the P.W. estimator is more efficient than the C.O. because of the extra degree of freedom. Since the Durbin estimator is more efficient than the C.O. for ρ larger than 0.3 and is not significantly less efficient for other values of ρ we definitely gain efficiency by using the Durbin $\hat{\rho}$ in the P.W. The relative efficiency of the P.W. with the Durbin estimated $\hat{\rho}$ compared to the Durbin is represented in Table 7 for selected values of ρ. Comparison of these results with Table 8 (which provides a direct comparison of the Durbin and the original P.W. estimators) shows that except for high λ the gain in efficiency is largely the result of the extra degree of freedom. Even though the Durbin $\hat{\rho}$ gives a "better" weight to the first observation this has only a small effect on the efficiency of the second stage.

The non-linear estimator is a maximum likelihood estimator under normality and convergence assumptions. The two-stage estimators are one-iteration methods while the non-linear estimator continues to iterate until convergence is achieved. In large samples both types of estimators are equally efficient. In principle, the non-linear procedure uses, and hence, also should provide more information than one iteration procedures in small samples. To verify this we computed the relative efficiency of the non-linear and the Durbin estimator in Table 9.

The non-linear estimator appears to be more efficient only when ρ and λ are both negative and very high. For the most part the non-linear estimator is about as efficient as the Durbin one, but it is certainly no improvement on it. In addition, a number of "outliers" give it very high M.S.E.'s for high ρ and λ of opposite sign. We examined several of these outliers in greater detail and found no multiple minima or other irregularities in the RSS function. What happens is that for high λ and ρ, the RSS function is often very flat over a significant range of parameter values. In these cases the non-linear procedure may converge to the wrong values of ρ and β. The point of convergence is not a local minimum, but the function is effectively stationary in its neighborhood. This type of "outliers" could be eliminated through the use of a scanning routine. But even when the non-linear procedure converged to the "right" values,

TABLE 8. RELATIVE EFFICIENCY OF THE P.W. ESTIMATOR COMPARED TO THE DURBIN ESTIMATOR OF β

(M.S.E.(Durbin)/M.S.E.(P.W.))

	LAMBDA									
RHO	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	0.99
0.99	0.87	0.75	0.88	0.62	0.78	0.64	0.69	0.47	0.72	0.48
0.9	1.04	1.18	0.98	0.61	0.83	1.00	0.85	0.64	0.56	0.61
0.8	0.98	1.01	0.84	0.83	0.79	0.89	0.84	0.82	1.00	1.05
0.7	1.00	1.07	0.96	1.05	0.90	0.94	0.87	1.04	0.88	1.15
0.6	0.97	1.06	0.91	0.92	0.95	0.89	0.82	1.16	1.10	1.14
0.5	1.09	1.13	0.92	0.87	0.94	1.10	1.01	1.07	0.85	0.95
0.4	1.04	1.05	0.88	1.03	1.00	1.13	1.07	1.21	1.09	1.08
0.3	1.02	1.19	0.99	1.03	0.99	1.14	1.03	1.24	1.09	1.16
0.2	1.01	1.07	0.97	0.98	1.05	0.89	0.89	1.10	1.23	1.00
0.1	1.11	1.05	1.06	0.90	1.14	1.05	1.06	1.01	1.14	1.32
0.0	1.10	1.32	0.99	1.07	0.98	1.21	1.05	1.04	1.04	0.81
-0.1	1.21	1.04	0.99	1.12	1.03	1.17	1.07	1.14	1.09	0.91
-0.2	0.96	1.24	0.99	1.02	1.12	1.05	1.03	1.24	1.14	1.37
-0.3	1.14	1.09	0.96	1.03	1.08	0.89	0.96	1.02	1.20	1.10
-0.4	1.26	1.02	1.07	1.00	1.13	1.09	1.00	1.06	1.11	1.47
-0.5	1.10	0.95	1.12	1.06	1.12	0.98	1.11	0.97	1.06	1.38
-0.6	1.27	0.92	0.96	1.14	0.97	0.97	1.07	1.12	1.15	1.67
-0.7	1.02	0.89	1.01	1.04	0.91	1.05	0.97	0.96	0.92	1.48
-0.8	0.91	0.94	0.72	0.95	1.01	1.06	0.97	0.95	1.00	1.40
-0.9	0.76	0.71	0.96	1.08	1.04	0.98	1.04	1.00	1.08	1.39

as it did most of the time, the resulting estimates were not appreciably better than those obtained from one-iteration methods. Thus there is little gain to be had from more complex procedures in samples of this size.[14]

One never knows, of course, the true ρ in practice. All one can have is an estimate of it, $\hat{\rho}$. Can one choose a better estimator on the basis of an estimated $\hat{\rho}$? To answer this question we investigated the performance of several mixed estimators based on the O.L.S. $\hat{\rho}$. These estimators switch to the P.W. method

[14] The P.W. estimator gained efficiency by making partial use of the first observation. In principle, the non-linear estimator should also be more efficient if we provided it with information on how the first observation was generated. To verify this we defined the residual sum of squares (RSS) function as follows:

$$\text{RSS} = \sum_{2}^{20} w_t^2 + (1 - \rho^2)u_1^2$$

and minimized it with respect to ρ and β by using the non-linear procedure. Except for ρ and λ very close to unity this showed some improvement over the original non-linear procedure. But the gain in efficiency was not large enough to change the earlier conclusions about the relative performance of non-linear and one-iteration methods.

TABLE 9. RELATIVE EFFICIENCY OF THE DURBIN AND NON-LINEAR ESTIMATORS OF β

[M.S.E.(non-linear)/M.S.E.(Durbin)]

	LAMBDA									
RHO	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	0.99
0.99	52.00	1.05	1.04	1.25	1.02	1.01	1.02	1.18	0.89	1.35
0.9	7.70	1.01	1.02	1.72	1.05	1.00	1.03	1.19	1.30	1.13
0.8	1.02	2.02	1.05	1.06	1.38	1.16	1.05	1.14	0.99	1.33
0.7	1.01	1.01	1.07	1.01	0.97	1.01	1.03	1.05	1.24	1.17
0.6	2.22	1.03	0.98	0.98	1.09	1.07	1.19	1.14	1.09	1.24
0.5	0.99	1.10	1.06	1.01	1.00	1.04	1.03	0.94	1.13	1.47
0.4	1.68	1.02	1.12	1.05	1.07	1.04	0.99	1.01	1.07	1.12
0.3	1.25	1.04	1.21	1.02	1.05	1.00	1.04	1.05	1.11	1.07
0.2	1.01	1.05	1.00	1.03	1.04	1.09	1.06	1.05	1.05	1.16
0.1	1.14	1.10	1.07	1.00	1.05	1.06	1.06	1.09	1.02	1.09
0.0	1.01	1.11	1.06	1.31	1.01	1.00	1.02	0.95	1.01	1.24
-0.1	1.02	1.19	1.04	0.97	1.09	1.00	0.97	1.09	1.02	1.30
-0.2	1.12	1.09	1.07	1.04	1.07	1.06	1.03	0.92	1.04	1.37
-0.3	1.18	1.03	1.03	1.03	0.98	1.12	1.04	1.74	0.98	4.70
-0.4	1.18	1.06	1.02	1.02	0.99	0.99	0.97	0.94	1.15	4.12
-0.5	1.08	1.01	0.95	1.00	0.95	0.95	1.02	0.99	0.98	3.08
-0.6	1.05	0.95	1.01	1.04	1.06	1.02	0.92	1.00	0.99	2.66
-0.7	0.99	0.94	0.98	0.92	1.08	1.00	1.00	1.01	2.93	6.15
-0.8	0.92	0.93	1.01	0.91	0.93	0.96	1.01	1.06	3.49	7.04
-0.9	0.67	0.89	0.98	0.99	1.00	0.98	1.00	1.00	1.00	23.21

if the computed $\hat{\rho}$ is higher than some critical value, and stay with the O.L.S. estimator otherwise. None of these estimators was superior to the others over the whole range of parameter values, but a mixed estimator switching to P.W. when $|\hat{\rho}| \geq .3$ appears to be a good compromise, losing very little efficiency to the "best" estimators over the whole range of parameters.[15] A comparison of this estimator to the P.W. one is presented in Table 10.

6. SUMMARY

We started this investigation with the hunch that the sampling variation in $\hat{\rho}$ may, in small samples, negate much of the promised gain from "efficient" procedures. This hunch proved to be wrong. For the sample size ($T=20$) and the type of x series examined (first order Markov with λ from $-.8$ to .99) there is a significant gain in efficiency to be had from using two stage estimation

[15] If one suspects that the true ρ is positive, one should base the mixed estimator on the less biased Durbin $\hat{\rho}$.

TABLE 10. THE RELATIVE EFFICIENCY OF A MIXED ESTIMATOR BASED ON THE SWITCHING POINT $|\hat{\rho}| \geq 0.3$

Efficiency = M.S.E.(P.W.)/M.S.E.(Mixed: $|\hat{\rho}_{\text{critical}}| = 0.3$)

	LAMBDA								
RHO	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8
0.9	1.00	1.00	1.00	0.94	0.95	0.98	0.98	0.99	0.98
0.8	0.95	0.97	1.00	0.96	0.89	0.91	1.01	0.95	0.98
0.7	1.00	1.00	0.91	1.04	0.88	0.95	0.88	0.91	0.89
0.6	0.92	0.99	0.98	0.95	0.97	0.99	0.97	0.94	0.97
0.5	0.97	0.97	0.80	0.80	0.94	0.90	0.96	0.97	0.97
0.4	0.93	0.98	0.98	0.99	0.90	0.92	1.01	1.06	1.02
0.3	0.99	0.95	0.94	0.90	0.89	1.01	0.95	0.99	0.97
0.2	1.04	1.03	0.95	0.89	1.08	0.94	1.02	1.04	1.02
0.1	1.09	1.06	0.96	0.95	1.05	1.02	1.00	1.04	0.96
0.0	0.92	1.08	1.04	1.04	1.01	0.97	0.98	1.04	1.02
-0.1	0.98	1.02	0.98	1.01	1.03	1.04	0.95	N.C.	0.99
-0.2	N.C.	0.97	1.08	0.99	1.07	1.01	0.89	1.05	1.00
-0.3	0.93	0.98	0.97	0.99	0.97	0.99	1.00	0.91	1.01
-0.4	0.80	0.97	1.00	0.97	1.00	0.98	0.90	0.93	1.04
-0.5	0.96	0.99	0.99	0.94	0.93	0.89	1.01	0.98	N.C.
-0.6	0.93	0.93	0.95	1.08	1.03	0.90	0.95	1.02	0.99
-0.7	0.97	0.97	0.95	0.90	0.99	0.98	1.01	0.82	1.00
-0.8	1.00	0.99	0.96	1.00	1.00	0.95	1.01	1.00	1.00
-0.9	0.92	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00

n.c. =not computed.

procedures for moderate and high levels of serial correlation in the residuals [$|\hat{\rho}| > .3$] and very little loss from using such methods even when the true ρ is small.[16]

Among the various efficient estimators examined we find that a two-stage estimator based on the Durbin $\hat{\rho}$ (the coefficient of y_{t-1} in the expanded equation) and incorporating the first observation with appropriate weight in the sum of squares to be minimized, is likely to do best over a wider range of parameters than any of the other estimators examined. The gain from including the beginning observation may be quite high if it is significantly different from the average.[17] Only if ρ is very high, does the added error in the first observation outweight the information contained therein.

Non-linear maximum likelihood procedures are no improvement over the

[16] These results are, of course, conditional on the specification of the model. In particular, the critical value of $\hat{\rho}$ is a decreasing function of sample size.

[17] The beginning observation should be included in the suggested way only if it can be assumed that the process generating the disturbances started a long time ago and has operated and continues to do so without breaks.

simpler two-stage procedures in samples of this size. If they are going to be used, they should be supplemented by a scanning routine to guard against convergence to stationary but not maximum points on the likelihood function.

We conclude that where computational costs are a consideration, a compromise mixed strategy of switching to a second-stage only if the estimated $|\hat{\rho}| \geq .3$ should do relatively well over the whole parameter range.[18]

REFERENCES

[1] Chipman, J. S. (1965). "The Problem of Testing for Serial Correlation: the Story of a Dilemma", University of Minnesota, mimeographed unpublished paper.

[2] Cochrane, D., and Orcutt, G. H. (1949). "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms", *Journal of American Statistical Association*, 44 (1), 32–61.

[3] Dhrymes, P. J. (1966). "On the Treatment of Certain Recurrent Non-linearities in Regression Analysis", *Southern Economic Journal*, Vol. 33.

[4] Durbin, J. (1960). "The Fitting of Time-Series Models", *Review of the International Statistical Institute*, 28, 233–243.

[5] Hartley, H. O. (1961). "The Modified Gauss-Newton Method for the Fitting of Non-linear Regression Functions by Least Squares", *Technometrics*, 3.

[6] Hildreth, C. (1966). "Asymptotic Distribution of Maximum Likelihood Estimators in Linear Models with Autoregressive Disturbances", RAND Memorandum, RM-5059-PR, Santa Monica.

[7] Hildreth, C. and Lu, J. Y. (1960). *Demand Relations with Autocorrelated Disturbances*, Res. Bull. 276, Michigan State AES.

[8] Johnston, J. (1963). *Econometric Methods*. New York, McGraw-Hill.

[9] Kendall, M. G. (1954). "Note on Bias in the Estimation of Autocorrelation", *Biometrika*, 41, 403–404.

[10] Malinvaud, E. (1966). *Statistical Methods of Econometrics*. Chicago: Rand McNally and Company.

[11] Orcutt, G. H. and Winokur, H. S. (1969). "First Order Autoregression: Inference, Estimation, and Prediction", *Econometrica*, forthcoming.

[12] Prais, S. J., and Winsten, C. B. (1954). "Trend Estimators and Serial Correlation," unpublished Cowles Commission discussion paper: Stat. No. 383, Chicago.

[13] Rao, P. (1968). "The Generalized Gauss-Newton Procedure to fit Nonlinear Regressions", unpublished paper, Chicago.

[14] Watson, G. S. (1955). "Serial Correlation in Regression Analysis", *Biometrika*, 42, 327–341.

[15] Watson, G. S., and Hannan, E. J. (1956). "Serial Correlation in Regression Analysis II", *Biometrika*, 43.

[16] Wold, H. (1949). "On Least Squares Regression with Autocorrelated Error Terms", *Bulletin of the International Statistical Institute*, 32 (2).

[17] Zellner, A., and Tiao, G. (1964). "Bayesian Analysis of the Regression Model with Autocorrelated Errors", *Journal of the American Statistical Association*, 59.

APPENDIX

Equation 1.

$$E\left(\sum_{1}^{T} x_t u_t\right)^2 \simeq \sigma_u^2 \cdot \sum_{1}^{T} x_t^2 \cdot \frac{1+\rho\lambda}{1-\rho\lambda}$$

[18] Among the problems not examined in this paper is the possibility of improving upon the performance of the two-stage methods by adjusting $\hat{\rho}$ upward for its known downward bias. A recent Monte-Carlo experiment by Orcutt and Winokur [11] indicates that this is not a promising avenue. The reduction of bias leads to an increase in variance with little or no improvement in the mean square error. Similarly, we have not explored the utilization of prior knowledge of the bounds of ρ via Bayesian techniques. On this see the important paper by Zellner and Tiao [17].

Proof:

$$\left(\sum_1^T x_t u_t\right)^2 = \sum_1^T \sum_1^T x_t x_{t'} u_t u_{t'}$$

$$E\left(\sum_1^T x_t u_t\right)^2 = \sigma_u^2 \cdot \sum x_t^2 + 2\rho\sigma_u^2(x_1x_2 + x_2x_3 + \cdots + x_{T-1}x_T)$$

$$+ 2\rho^2\sigma_u^2(x_1x_3 + x_2x_4 + \cdots + x_{T-2}x_T) + \cdots.$$

Neglecting the cross product terms of x and v for large samples,

$$E\left(\sum_1^T x_t u_t\right)^2 = \sigma_u^2 \sum_1^T x_t^2 + 2\rho\lambda\sigma_u^2 \sum_1^{T-1} x_t^2 + 2\rho^2\lambda^2\sigma_u^2 \sum_1^{T-2} x_t^2 + \cdots$$

$$= \sigma_u^2 \sum_1^T x_t^2 \left[1 + 2\rho\lambda + 2\rho^2\lambda^2 + 2\rho^3\lambda^3 + \cdots\right.$$

$$\left. - 2\rho\lambda \frac{x_T^2}{\sum_1^T x_t^2} - 2\rho^2\lambda^2 \frac{x_T^2 + x_{T-1}^2}{\sum_1^T x_t^2} - \cdots\right]$$

Since ρ and λ are less than one in magnitude, for reasonably large samples, we may ignore the negative terms in the above expression. The series is convergent. Hence

$$E\left(\sum_1^T x_t u_t\right)^2 \simeq \sigma_u^2 \cdot \sum_1^T x_t^2 \left[1 + \frac{2\rho\lambda}{1 - \rho\lambda}\right]$$

$$\simeq \sigma_u^2 \cdot \sum_1^T x_t^2 \cdot \frac{1 + \rho\lambda}{1 - \rho\lambda}.$$

Equation 2.

$$E\left(\sum_2^T e_t e_{t-1}\right) \simeq \sigma_u^2 \left[\rho\left(T - 1 - \frac{1 + \rho\lambda}{1 - \rho\lambda}\right) - (\rho + \lambda)\right]$$

Proof:

$$y_t = \hat{\beta} x_t + e_t$$

where

$$\hat{\beta} = \sum_1^T x_t y_t \Big/ \sum_1^T x_t^2$$

$$e_t = y_t - x_t \cdot \frac{\sum_1^T x_t y_t}{\sum_1^T x_t^2}$$

$$e_t = \beta x_t + u_t - x_t \cdot \frac{\Sigma x(\beta x_t + u_t)}{\Sigma x_t^2}$$

$$= u_t - \frac{x_t}{\sum_1^T x_t^2}\left(\sum_1^T x_t u_t\right)$$

$$e_t e_{t-1} = \left[u_t - \frac{x_t}{\sum_1^T x_t^2}(\Sigma x_t u_t)\right]\left[u_{t-1} - \frac{x_{t-1}}{\sum_1^T x_t^2}\sum_1^T x_t u_t\right]$$

$$= u_t u_{t-1} - \frac{x_t u_{t-1}}{\Sigma x_t^2}\cdot\Sigma x_t u_t - \frac{x_{t-1}u_t}{\Sigma x_t^2}\Sigma x_t u_t + \frac{x_t x_{t-1}}{\Sigma x_t^2\cdot\Sigma x_t^2}(\Sigma x_t u_t)^2$$

$$\Sigma e_t e_{t-1} = \Sigma u_t u_{t-1} - \frac{\Sigma x_t u_{t-1}\cdot\Sigma x_t u_t}{\Sigma x_t^2} - \frac{\Sigma x_{t-1}u_t \Sigma x_t u_t}{\Sigma x_t^2} + \frac{\Sigma x_t x_{t-1}}{\Sigma x_t^2\cdot\Sigma x_t^2}\cdot(\Sigma x_t u_t)^2$$

$\Sigma x_t u_{t-1}\cdot\Sigma x_t u_t$

$$= x_2u_1(x_1u_1 + x_2u_2 + \cdots + x_Tu_T)$$
$$+ x_3u_2(x_1u_1 + x_2u_2 + \cdots + x_Tu_T)$$
$$+ \cdots\cdots\cdots\cdots\cdots\cdots$$

$$E\left(\frac{\Sigma x_t u_{t-1}\Sigma x_t u_t}{\Sigma x_t^2}\right)$$

$$= \frac{\sigma_u^2}{\Sigma x_t^2}\cdot\begin{bmatrix} x_2x_1 + x_2^2\rho + x_2x_3\rho^2 + \cdots \\ + x_3x_1\rho + x_3x_2 + x_3^2\rho + \cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots \end{bmatrix}$$

Similarly

$$E\left(\frac{\Sigma x_{t-1}u_t\Sigma x_t u_t}{\Sigma x_t^2}\right) = \frac{\sigma_u^2}{\Sigma x_t^2}\cdot\begin{pmatrix} x_1^2\rho + x_1x_2 + x_1x_3\rho + \cdots \\ + x_2x_1\rho^2 + x_2^2\rho + x_2x_3\rho^2 + \cdots \end{pmatrix}$$

Since ρ and λ are smaller than 1 in magnitude

$$E\left(\frac{\Sigma x_t u_{t-1}\Sigma x_t u_t}{\Sigma x_t^2} + \frac{\Sigma x_{t-1}u_t\Sigma x_t u_t}{\Sigma x_t^2}\right)$$

$$= 2\sigma_u^2(\rho + \lambda\rho^2 + \lambda^2\rho^3 + \cdots + \lambda + \lambda^2\rho + \lambda^3\rho^2 + \cdots)$$
$$= 2\sigma_u^2\cdot\left(\rho + \lambda + \frac{\rho^2\lambda}{1-\rho\lambda} + \frac{\lambda^2\rho}{1-\rho\lambda}\right)$$
$$= 2\sigma_u^2\cdot\frac{\rho+\lambda}{1-\rho\lambda}$$

using Equation 1.

$$E\left(\sum_{1}^{T} e_t e_{t-1}\right)$$

$$\simeq (T-1)\rho\sigma_u^2 + \lambda\sigma_u^2 \cdot \frac{1+\rho\lambda}{1-\rho\lambda} - 2\sigma_u^2 \frac{\rho+\lambda}{1-\rho\lambda}$$

$$= \sigma_u^2\left((T-1)\rho + \frac{\lambda(1+\rho\lambda) - 2(\rho+\lambda)}{1-\rho\lambda}\right)$$

$$= \sigma_u^2\left[(T-1)\,\rho + \frac{\rho(1+\rho\lambda) - \rho(1+\rho\lambda) + \lambda(1+\rho\lambda) - 2(\rho+\lambda)}{1-\rho\lambda}\right]$$

$$= \sigma_u^2\left[\rho\left(T-1-\frac{1+\rho\lambda}{1-\rho\lambda}\right) - (\rho+\lambda)\right]$$

Equation 3.

$$E\left(\sum_{2}^{T} e_t^2\right) \simeq \sigma_u^2\left(T-1-\frac{1+\rho\lambda}{1-\rho\lambda}\right)$$

Proof:

$$e_t = u_t - \frac{x_t}{\sum_{1}^{T} x_t^2}\sum_{1}^{T} x_t u_t$$

$$e_t^2 = u_t^2 + \frac{x_t^2}{\sum_{1}^{T} x_t^2 \sum_{1}^{T} x_t^2}\left(\sum_{1}^{T} x_t u_t\right)^2 - 2\frac{x_t u_t}{\Sigma x_t^2}\cdot\sum_{1}^{T} x_t u_t$$

Using equation 1 of the Appendix,

$$E\left(\sum_{2}^{T} e_t^2\right) \simeq (T-1)\sigma_u^2 + \sigma_u^2 \cdot \frac{1+\rho\lambda}{1-\rho\lambda} - 2\sigma_u^2 \frac{1+\rho\lambda}{1-\rho\lambda}$$

$$\simeq (T-1)\sigma_u^2 - \sigma_u^2 \cdot \frac{1+\rho\lambda}{1-\rho\lambda}$$

$$\simeq \sigma_u^2\left[(T-1) - \frac{1+\rho\lambda}{1-\rho\lambda}\right]$$