A Maximum Likelihood Procedure for Regression With Autocorrelated Errors

Article *in* Econometrica · January 1978

DOI: 10.2307/1913644 · Source: RePEc

2 authors:

James G. Mackinnon
Queen's University

Charles M. Beach
Queen's University

Econometrica, Vol. 46, No. 1 (January, 1978)

A MAXIMUM LIKELIHOOD PROCEDURE FOR REGRESSION WITH AUTOCORRELATED ERRORS

BY CHARLES M. BEACH AND JAMES G. MACKINNON[1]

The widely used Cochrane–Orcutt and Hildreth–Lu procedures for estimating the parameters of a linear regression model with first-order autocorrelation typically ignore the first observation. An alternative maximum likelihood procedure which incorporates the first observation and the stationarity condition of the error process is proposed in this paper. It is similar to the Cochrane–Orcutt procedure, and appears to be at least as computationally efficient. This estimator is superior to the conventional ones on theoretical grounds, and sampling experiments suggest that it may yield substantially better estimates in some circumstances.

1. INTRODUCTION

IT HAS BEEN KNOWN for some time that alternative treatments of the initial observations of an autoregressive error model can lead to different results when a generalized least squares or Aitken [**1**] estimation procedure is used. Consider the standard single-equation linear regression model where the error terms follow a first-order autoregressive process,

$$y = X\beta + u, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \operatorname{var} \varepsilon_t = \sigma^2 \qquad (t = 1, \ldots, T). \tag{1}$$

Least squares estimates are typically based on the Cochrane–Orcutt transformation (see [**3**]),

$$P = \begin{bmatrix} -\rho & 1 & 0 & \ldots & & & \\ 0 & -\rho & 1 & 0 & \ldots & & \\ & & & \ldots & 0 & -\rho & 1 \end{bmatrix}$$

which ignores the information contained in $\hat{u}_1$. If the error process is stationary, the variance of u_1 is $\sigma^2/(1-\rho^2)$. Thus, as Prais and Winsten [**9**] and others have shown, the appropriate transformation is

$$Q = \begin{bmatrix} (1-\rho^2)^{\frac{1}{2}} & 0 & \ldots & & & \\ -\rho & 1 & 0 & \ldots & & \\ & & \ldots & 0 & -\rho & 1 \end{bmatrix}$$

which retains the first observation with a weight of $(1-\rho^2)^{\frac{1}{2}}$. Gurland [**5**], Kadiyala [**7**], and Rao and Griliches [**10**] have studied the relative efficiency of GLS estimators based on these two transformations, and found that using P instead of Q can cause a substantial loss of efficiency.

None of the procedures examined in the studies to which we just referred is maximum likelihood when ρ has to be estimated. The ML procedures which are typically used are those of Cochrane and Orcutt [**3**] and Hildreth and Lu [**6**]. Both

[1] We are indebted to Richard E. Quandt, the members of the Quantitative Workshop at Queen's University, and particularly an anonymous referee, for comments on an earlier draft.

of these are least-squares procedures using P, and are thus equivalent to maximizing the likelihood function for the last $T-1$ observations. Asymptotically, disregarding the first observation makes no difference; but in small samples, it may make a substantial difference, which is what the sampling experiments mentioned suggest. It would therefore seem desirable to employ an ML procedure that does not disregard the first observation and at the same time incorporates the *a priori* stationarity restriction that $|\rho|<1$. In this paper, such a procedure is developed. It is similar to the Cochrane–Orcutt technique, and appears to be at least as fast to compute. We also present some Monte Carlo results which suggest that, in many cases, our procedure yields substantially more efficient estimates than the conventional techniques.

2. DERIVATION OF THE LIKELIHOOD FUNCTION

The likelihood function for the spherical normal error vector ε is (up to a factor or proportionality)

$$L(\varepsilon) \propto \sigma^{-T} \exp\left[-\varepsilon'\varepsilon/2\sigma^2\right].$$

If ε is expressed as Qu, $L(\varepsilon)$ is transformed into

$$L(u) \propto \sigma^{-T}(1-\rho^2)^{\frac{1}{2}} \exp\left[-u'Q'Qu/2\sigma^2\right]$$

where $(1-\rho^2)^{\frac{1}{2}}$ is the Jacobian of the transformation. Concentrating out σ^2 yields the concentrated log-likelihood function

$$\mathscr{L} = \text{const.} \frac{1}{2}\log(1-\rho^2) - \frac{T}{2}\log\left[(1-\rho^2)(y_1 - X_1\beta)^2 + \sum_{t=2}^{T}(y_t - X_t\beta - \rho y_{t-1} + \rho X_{t-1}\beta)^2\right]. \tag{2}$$

It can be seen that ML estimates based on (2) differ from conventional ones because of the presence of two extra terms in $\mathscr{L}$: $(1-\rho^2)(y_1-X_1\beta)^2$ in the sum of squares expression, and $1/2\log(1-\rho^2)$. The first of these alone ensures that the initial observation will have some effect upon the estimates, as in the GLS procedures. The second, which is not taken into account in the GLS procedures, explicitly constrains the stationarity condition to hold. On purely theoretical grounds, then, one should strongly prefer the present full maximum likelihood (or FML) procedure to conventional ones which disregard the first observation and associated stationarity condition, and to two-step GLS procedures.

3. A MAXIMIZATION PROCEDURE

Many procedures can be used for maximizing the likelihood function (2). The estimate of β which maximizes (2) conditional on ρ is simply

$$\tilde{\beta} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \quad \text{where} \quad X^* = QX \quad \text{and} \quad y^* = Qy. \tag{3}$$

Thus one can conduct a one-dimensional search over ρ on the open interval $(-1, 1)$, calculating $\tilde{\beta}$ at each step and evaluating (2) instead of the residual sum of squares. The only problem with such a procedure is that it is relatively costly: scanning at intervals of .1 requires 19 least squares computations (and yields a rather inaccurate estimate of ρ), while in our experience more sophisticated golden section search to an accuracy of five digits requires about 35 least squares computations. For maximizing the conventional likelihood function, the Cochrane–Orcutt technique is usually much less costly than such procedures. We have therefore developed an analogous technique for maximizing the full likelihood function.

The technique proceeds by alternately maximizing (2) with respect to β, ρ held fixed, and maximizing (2) with respect to ρ, β held fixed. This usually begins with ρ equal to zero, and ends when two successive values of ρ are sufficiently close. Two separate maximizations must be performed at each iteration. The solution to the first of these is (3), and we now develop the solution to the second. Differentiating (2) with respect to ρ, setting equal to zero, and rearranging terms yield the first-order condition obtained by Koopmans [**8**] and Anderson [**2**, p. 354]:

(4) $$f(\rho) \equiv \rho^3 + a\rho^2 + b\rho + c = 0$$

where

$$a = -(T-2)\sum A_tA_{t-1}/[(T-1)(\sum A_{t-1}^2 - A_1^2)],$$
$$b = [(T-1)A_1^2 - T\sum A_{t-1}^2 - \sum A_t^2]/[(T-1)(\sum A_{t-1}^2 - A_1^2)],$$
$$c = T\sum A_tA_{t-1}/[(T-1)(\sum A_{t-1}^2 - A_1^2)],$$

and $A_t = y_t - X_t\beta$ for given β; the summations run from $t = 2$ to T. As Anderson shows, $f(-1) > 0$, $f(+1) < 0$, and $f(0)$ has the same sign as $\Sigma\, A_tA_{t-1}$. Consequently, (4) has a unique real root in $(-1, 0)$ when $\Sigma\, A_tA_{t-1} < 0$, a root of zero when $\Sigma\, A_tA_{t-1} = 0$, and a unique real root in $(0, 1)$ when $\Sigma\, A_tA_{t-1} > 0$.

The cubic $f(\rho)$ belongs to the "irreducible case" of cubic equations, and the solution of (4) follows Uspensky [**12**, 84–93]. Define $p = b - a^2/3$ and $q = c - ab/3 + 2a^3/27$. Then calculate ϕ between zero and π radians as

$$\phi = \cos^{-1}\left((q\sqrt{27})/(2p\sqrt{-p})\right).$$

The desired root of (4) is

(5) $$\tilde{\rho} = -2\sqrt{-p/3}\cos(\phi/3 + \pi/3) - a/3.$$

Equations (3) and (5) thus provide the two main components of an algorithm for maximizing (2). Since the value of the likelihood function increases at every step, the algorithm must eventually get arbitrarily close to a relative maximum, although, as in the case of the Cochrane–Orcutt procedure, one cannot be certain that this is also a global maximum (see Sargan [**11**]). In our experience, this algorithm works extremely rapidly, requiring, on average, between four and seven least squares computations to achieve better than five digit accuracy.[2]

[2] When programming this procedure, we checked that it yielded the same answers as direct maximization of (2) by golden search. The answers always agreed to at least five digits which was the termination tolerance for both methods.

As with standard maximum likelihood procedures, estimated asymptotic standard errors may be found from the formula for the Cramer–Rao lower bounds. In the case of (2), these expressions can be derived analytically as done in the Appendix.

4. MONTE CARLO RESULTS

In order to find out how the FML estimates differ from the conventional ones, we conducted a sampling experiment. The model examined was

$$Y_t = \beta_1 + \beta_2 X_t + u_t, \qquad u_t = \rho u_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \text{NID}(0, .0036).$$

Initially, the X's were chosen to contain a large trend component, as a realization of

$$X_t = \exp(.04t) + w_t, \qquad w_t \sim \text{NID}(0, .0009).$$

Based on the results of Rao and Griliches [**10**], it was expected that gains from the full procedure would be substantial in such a case. The true values of β_1 and β_2 were both one. We present results for two different sample sizes, 20 and 50, and three different values of ρ, .6, .8, and .99. Each experiment involved two hundred replications. On each replication, both the conventional and full maximum likelihood estimates were computed for a given realization of the ε's,[3] using the Cochrane–Orcutt procedure and the FML procedure described above.

Table I presents biases and root mean square errors for the experiments just described. The first element in each cell refers to the full procedure, the second to the conventional one. In every case, there are gains to be had (in terms of RMSE's) from using the full procedure. For ρ, these gains are quite small; but for β_2, they are always substantial; and for β_1, they are often dramatic. Table I also records the number of replications for which the full estimates were closer to the true values than the conventional estimates; an "*" indicates that this number is significantly different from 100 at the .05 level. The FML estimates of β_1 and β_2 always do better than the conventional ones on this criterion, and the difference is usually significant; but their superiority is not as dramatic as it is in terms of RMSE's. The reason for this is that the conventional procedure seems to yield extremely poor estimates occasionally, generally when $\tilde{\rho}$ is close to unity, and this results in large RMSE's.

We believe the trending X case is a realistic one. However, we also considered a case where the X's were NID(0, .0625), using a sample size of 20 and the same values of ρ as before. The results of these experiments are also presented in Table I. There is now virtually no difference between the FML and conventional estimates of β_2; but the full procedure provides dramatically better estimates of β_1.

The relative gains from using the full procedure are highlighted in Table II, which tabulates the ratio of the RMSE from the conventional procedure to the RMSE from the FML procedure. This ratio is always higher for the intercept, β_1,

[3] To ensure that the error series was stationary, u_1 was generated as $(1/(1-\rho^2))^{\frac{1}{2}}\varepsilon_1$.

TABLE I

BIAS AND ROOT MEAN SQUARE ERROR OF FML AND CONVENTIONAL PROCEDURES[a]

Parameter	True ρ	Trending X, $T=20$ Bias	RMSE	No.	Trending X, $T=50$ Bias	RMSE	No.	Nontrending X, $T=20$ Bias	RMSE	No.
ρ	.6	−.250	.328	96	−.075	.143	100	−.125	.249	94
		−.251	.335		−.076	.143		−.132	.252	
	.8	−.320	.386	104	−.099	.148	91	−.163	.263	96
		−.326	.396		−.102	.150		−.174	.272	
	.99	−.442	.497	104	−.153	.178	97	−.219	.281	97
		−.460	.514		−.161	.189		−.234	.300	
β_1	.6	.010	.138	121*	.002	.043	119*	−.001	.035	115*
		.072	.551		.000	.049		.000	.041	
	.8	.009	.215	123*	.005	.076	121*	−.002	.062	126*
		.086	.578		.002	.098		−.047	.495	
	.99	−.025	.597	110	.053	.436	127*	−.018	.450	109
		.047	1.283		.056	1.682		.107	3.078	
β_2	.6	−.006	.088	125*	−.001	.011	106	.002	.041	101
		−.021	.134		.000	.012		.002	.041	
	.8	−.006	.134	119*	−.001	.018	109	.001	.037	98
		−.024	.183		−.001	.022		.001	.037	
	.99	.005	.235	113	−.005	.059	111	.001	.033	107
		−.011	.266		−.004	.065		.001	.033	

[a] The first element in each cell refers to the FML procedure, and the second refers to the conventional ML procedure.

TABLE II
RATIOS OF RMSE'S

Parameter	True ρ	Trending X, $T=20$	Trending X, $T=50$	Nontrending X, $T=20$
ρ	.6	1.0209	1.0001	1.0129
	.8	1.0260	1.0153	1.0352
	.99	1.0329	1.0613	1.0681
β_1	.6	3.9797	1.1385	1.1919
	.8	2.6894	1.2871	7.9537
	.99	2.1498	3.8572	6.8382
β_2	.6	1.5130	1.0925	1.0011
	.8	1.3609	1.1790	1.0031
	.99	1.1358	1.1126	.9994

than for the slope coefficient, β_2. We had expected the gains from employing the full procedure to be much smaller for $T=50$ than for $T=20$, but that is rarely the case. Thus it may require a very large sample indeed for the asymptotic equivalence of the conventional and full procedures to render the two equally efficient. As expected, the estimates of ρ are always biased downwards. In every case, this bias is significant at the .01 level, using a binomial test on the number of times the estimate fell below the true value. Using a similar test, the estimates of β_1 and β_2 are never significantly biased at the .05 level.

One interesting and unexpected result of these experiments is the discovery that the technique developed in this paper may well be computationally less expensive than the Cochrane–Orcutt technique. One iteration of the FML procedure costs slightly more than one iteration of Cochrane–Orcutt essentially because the solution for $\tilde{\rho}$ is more complicated. But our experience is that the former procedure often requires substantially fewer iterations than Cochrane–Orcutt; the average number of least squares regressions that each required is presented in Table III. In both cases, ρ was initially set equal to zero, and the algorithm terminated when $\tilde{\rho}$ changed by less than 10^{-5} between two successive iterations.

TABLE III
AVERAGE NUMBER OF OLS ITERATIONS REQUIRED

	Full Procedure	Conventional Procedure
Trending X:		
$T=20, \rho=.6$	5.90	9.31
$\rho=.8$	6.55	12.21
$\rho=.99$	7.24	14.42
$T=50, \rho=.6$	4.66	4.79
$\rho=.8$	5.14	5.59
$\rho=.99$	5.81	12.74
Nontrending X:		
$T=20, \rho=.6$	6.03	6.53
$\rho=.8$	5.92	11.31
$\rho=.99$	5.84	38.86

5. CONCLUSION

It has been argued that maximum likelihood estimators of regression models with first-order serial correlation should not disregard the first observation. The full procedure has the very important theoretical advantage that the stationarity condition $|\rho| < 1$ is incorporated *a priori* in the likelihood function itself. A computationally efficient technique has been developed for maximizing the full likelihood function in the linear regression case. It is no more expensive than the Cochrane-Orcutt technique, and may often be substantially less expensive. The results of sampling experiments suggest that this full procedure may yield substantially better estimates than conventional procedures. We therefore conclude that future empirical work employing the maximum likelihood approach should not ignore the initial observation and the associated stationarity condition.

Queen's University

Manuscript received February, 1976; final revision received December, 1976.

APPENDIX

An estimate of the asymptotic variance-covariance matrix for $\tilde{\theta}' = (\tilde{\sigma}^2, \tilde{\rho}, \tilde{\beta}')$ may be found by taking the inverse of minus the expected value of the matrix of second derivatives of the log-likelihood function with respect to $\tilde{\theta}$. Provided that all of the X's are exogenous, this matrix is

$$\begin{bmatrix} A & C & 0 \\ C & B & 0 \\ 0 & 0 & D \end{bmatrix}^{-1}$$

where

$$A = T/2\tilde{\sigma}^4,$$

$$B = (1+\tilde{\rho}^2)/(1-\tilde{\rho}^2)^2 + (1/\tilde{\sigma}^2)\left(\sum_{t=3}^{T} (y_{t-1} - X_{t-1}\tilde{\beta})^2\right),$$

$$C = (1/\tilde{\sigma}^2)(\tilde{\rho}/(1-\tilde{\rho}^2)),$$

and the ijth element of D is equal to

$$(1/\tilde{\sigma}^2)\left[(1-\tilde{\rho}^2)X_{1i}X_{1j} + \sum_{t=2}^{T} (X_{ti} - \tilde{\rho}X_{t-1,i})(X_{tj} - \tilde{\rho}X_{t-1,j})\right].$$

When the X's contain a lagged dependent variable, the asymptotic covariance matrix can be modified following Cooper [**4**].

REFERENCES

[**1**] AITKEN, A. C.: "On Least Squares and Linear Combinations of Observations," *Proceedings of the Royal Society Edinburgh*, 55 (1934), 42–48.
[**2**] ANDERSON, T. W.: *The Statistical Analysis of Time Series*. New York: Wiley, 1971.
[**3**] COCHRANE, D., AND G. H. ORCUTT: "Applications of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44 (1949), 32–61.

[4] COOPER, J. PHILLIP: "Asymptotic Covariance Matrix of Procedures for Linear Regression in the Presence of First-Order Autoregressive Disturbances," *Econometrica*, 40 (1972), 305–310.

[5] GURLAND, J.: "An Example of Autocorrelated Disturbances in Linear Regression," *Econometrica*, 22 (1954), 218–227.

[6] HILDRETH, C., AND J. Y. LU: "Demand Relations with Autocorrelated Disturbances," Research Bulletin 276, Michigan State University Agricultural Experiment Station, 1960.

[7] KADIYALA, K. R.: "A Transformation Used to Circumvent the Problem of Autocorrelation," *Econometrica*, 36 (1968), 93–96.

[8] KOOPMANS, TJALLING: "Serial Correlation and Quadratic Forms in Normal Variables," *The Annals of Mathematical Statistics*, 13 (1942), 14–33.

[9] PRAIS, G. J., AND C. B. WINSTEN: "Trend Estimators and Serial Correlation," Cowles Commission Discussion Paper No. 383, Chicago, 1954.

[10] RAO, P., AND Z. GRILICHES: "Small-Sample Properties of Several Two-Stage Regression Methods in the Context of Auto-Correlated Errors," *Journal of the American Statistical Association*, 64 (1969), 253–272.

[11] SARGAN, J. D.: "Wages and Prices in the United Kingdom: A Study in Econometric Methodology," in *Econometric Analysis for National Economic Planning*, ed. by P. E. Hart, G. Mills, and J. K. Whitaker. London: Butterworth and Co. Ltd., 1964, pp. 25–54.

[12] USPENSKY, J. V.: *Theory of Equations*. New York: McGraw-Hill Book Co., 1948.