INTERNATIONAL ECONOMIC REVIEW
Vol. 14, No. 2, June, 1973

AN ADAPTIVE REGRESSION MODEL*

BY THOMAS F. COOLEY AND EDWARD C. PRESCOTT[1]

ECONOMETRICIANS frequently approximate complex behavioral and technological relationships using equations that are linear in a small number of unknown parameters. The effect of omitted variables, aggregation errors, and other errors in specification are included in the additive disturbance which is assumed, among other things, to be temporally uncorrelated. Utilizing time series data, linear regression analysis is then used to estimate parameters. The adaptive regression model developed in this paper would be used in the same manner, but it does not assume the disturbances are independent. Instead, it assumes the disturbances are the sum of not only a transitory element that has effect in the current period but also a permanent component whose effect persists into the future. If for example, omitted variables are subject to permanent and transitory changes, as is sometimes assumed in economic theory [**4**] and by the widely used adaptive forecasting model [**6**], these disturbances will have both permanent and transitory components. In the adaptive regression model the transitory disturbance can be thought of as the usual additive error term, while the permanent component causes random changes in the intercept value.

It is common practice in econometric research to test for serial correlation in the residuals. If the test indicates serial correlation is present it is typically assumed that the disturbances are subject to a first order auto-regressive process. In fact, such processes are likely to describe the true distribution of the disturbances only in rare instances. An auto-regressive error process implies that the effects of omitted factors *all* decay exponentially with time and at the *same* rate. This is an unreasonable assumption for most economic applications. Some omitted factors, such as labor union strikes or the vagaries of the weather, will have only transitory effects while other factors, like changes in tastes or technological developments, will have effects which persist into the future without decay. The auto-regressive assumption is often justified by the argument that omitted varibles are subject to an auto-regressive process. This argument holds, however, only if *all* omitted factors contributing to the additive disturbance are subject to auto-regressive processes with the *same* parameter. The widespread use of the auto-regressive correction in econometrics is explained by the fact that it accounts for serial correlation and is computationally efficient. The adaptive regression also explains serial correlation, is computationally efficient, and assumes an error structure which, in many situations, provides a better approximation of reality.

Of particular methodological interest is the proof of consistency of the maximum likelihood estimator for a subset of the unknown parameters when the

* Manuscript received August 16, 1972; revised January 5, 1973.
[1] The authors acknowledge helpful comments of Professors F. Gerard Adams, Michael D. McCarthy and Melvin Hinich.

observations are not identically and independently distributed and a consistent estimator does not exist for the entire parameter set.[2]

2. THE MODEL

The assumed structure is

$$y = \beta_{0t} + X_t'\beta^* + u_t \tag{1.1}$$

where y_t is the t-th observation of the dependent variable, β_{0t} the random intercept parameter for period t, x_t a $(k-1)$ component vector of predetermined explanatory variables, β^* a vector of unknown slope coefficients, and u_t the additive transitory disturbance. The elements β_{0t} are subject to permanent changes v_t:

$$\beta_{0,t+1} = \beta_{0,t} + v_t\,. \tag{1.2}$$

The u_t and v_t are all independent normal variates with mean 0 and variances

$$\text{var}\,(u_t) = (1-\gamma)\sigma^2 \text{ and } \text{var}\,(v_t) = \gamma\sigma^2\,, \tag{1.3}$$

with $0 \leq \gamma \leq 1$. The unknown parameter γ measures the relative importance of the permanent component, the larger its value the greater the importance of permanent change.

If $\gamma = 0$, the above structure is the multiple regression model with an unchanging intercept. On the other hand, if β^* equals the zero vector, this system reduces to the adaptive forecasting structure of **[6]**. Thus, the model is a generalization of both regression analysis and adaptive forecasting.[3]

The process generating the intercepts is not stationary and writing down the likelihood function is impossible. The likelihood function conditional on the value of the process at some point in time, however, is well defined for any finite set of these elements. One approach is to treat $\beta_{0,0}$ as an unknown parameter and the other $\beta_{0,t}$ as realizations of the random process, but, a more convenient selection for forecasting is the value of the intercept one period subsequent to the sample, $\beta_{0,T+1}$. Defining β_0 to be $\beta_{0,T+1}$ and using (1.3)

$$\beta_{0,t} = \beta_0 - \sum_{s=t}^{T} v_s\,. \tag{1.4}$$

Substituting for $\beta_{0,t}$ in (1.1) yields

$$y_t = \beta_0 + x_t'\beta^* + u_t - \sum_{s=t}^{T} v_s\,. \tag{1.5}$$

Except for the the dependence of the error terms, this is the usual linear regres-

[2] In a subsequent paper [**2**], we compared the predictive and estimation efficiencies of adaptive and conventional regression analysis with and without the auto-regressive correction. A Fortran Program for adaptive regression is available upon request.

[3] The structure is related to those considered by [**1**] and [**7**] from a Bayesian point of view. Our estimation procedure is computationally more efficient and we are able to develop asymptotic properties of the estimators.

sion structure.

To simplify notation let y be the T component vector of the y_t, β the k component vector

$$\beta' = [\beta_0, \beta_1^*, \beta_2^*, \cdots, \beta_{k-1}^*], \tag{1.6}$$

and X the $T \times k$ matrix

$$X = \begin{bmatrix} 1 & x_{11} & \cdot & \cdot & \cdot & x_{k-1,1} \\ 1 & x_{12} & \cdot & \cdot & \cdot & x_{k-1,2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1T} & \cdot & \cdot & \cdot & x_{k-1,T} \end{bmatrix}. \tag{1.7}$$

Define the $T \times T$ matrix Q_γ to be

$$Q_\gamma = (1 - \gamma)I + \gamma R, \tag{1.8}$$

where the $T \times T$ matrix R has (i,j)-th element

$$r_{ij} = \min [T - i + 1, T - j + 1]. \tag{1.9}$$

With this notation along with (1.5), it is easily verified that

$$y \sim N[X\beta, \sigma^2 Q_\gamma]. \tag{1.10}$$

If γ were known, Q_γ would be known and estimation would be a simple application of Aitken's generalized least squares analysis. The maximum likelihood estimator of β would be

$$B(\gamma) = (X'Q_\gamma^{-1}X)^{-1}XQ_\gamma^{-1}y, \tag{1.11}$$

and of σ^2

$$s^2(\gamma) = T^{-1}[y - XB(\gamma)]'Q_\gamma^{-1}[y - XB(\gamma)]. \tag{1.12}$$

Since γ is not known, search techniques must be utilized to determine the parameter set with maximum likelihood. The log likelihood function of the observations is (except for a constant)

$$L(y; \beta, \sigma^2, \gamma, X) = -\frac{T}{2}\ln \sigma^2 - \frac{1}{2}\ln |Q_\gamma| - \frac{1}{2\sigma^2}(y - X\beta)'Q_\gamma^{-1}(y - X\beta). \tag{1.13}$$

Inserting the conditional maximum likelihood estimators of β and σ^2 yields the concentrated likelihood function (except for a constant)

$$L_c(y; \gamma) = -\frac{T}{2}\ln s^2(\gamma) - \frac{1}{2}\ln |Q_\gamma|. \tag{1.14}$$

The estimation strategy is to search over the interval $0 \leq \gamma \leq 1$ and choose as the estimator of γ that value, say g, such that

$$L_c(y; g) \geq L_c(y; \gamma) \text{ all } \gamma \in [0, 1]. \tag{1.15}$$

The corresponding maximum likelihood estimates of β and σ^2 are then $B(g)$ and $s^2(g)$ respectively.

The above procedure is straight-forward but excessively expensive, given current computer technology, because the $T \times T$ matrix Q_γ must be inverted for each value of γ that is searched in the interval [0, 1]. Fortunately, the variables can be transformed so the covariance matrix is diagonal and the transformation does *not* depend upon γ.

Let P be the matrix whose rows are the set of orthonormal eigenvectors for R. Then $Py \sim N[PX\beta, \sigma^2 P[(1-\gamma)I + \gamma R]P']$ or

$$Py \sim N[PX\beta, \sigma^2 D(\gamma)] \tag{1.16}$$

where $D(\gamma)$ is diagonal. Letting r_i be the eigenvalue corresponding to the i-th row of P and d_i the (i, i)-th element of $D(\gamma)$,

$$d_i(\gamma) = (1-\gamma) + \gamma r_i. \tag{1.17}$$

The analytic expressions for the r_i and p_{ij} are given by the following:

Result

$$r_i = [2 + 2\cos\{2\pi(T-i+1)/(2T+1)\}]^{-1} \tag{1.18}$$

and

$$p_{ij} = (-1)^i 2(2T+1)^{-1/2} \sin[2\pi(T-i+1)(T-j+1)/(2T+1)]. \tag{1.19}$$

PROOF. The inverse of R is a tridiagonal matrix with 2's down the main diagonal except for a 1 in the first position, and -1 for the elements one position off the main diagonal. It is readily verified that the i-th row of P is an eigenvector corresponding to r_i^{-1} of the matrix R^{-1}. But, the eigenvectors of a symmetric matrix are the same as those of the inverse and the eigenvalues are the reciprocals. Thus, the rows of P are a set of orthonormal eigenvectors of R and the eigenvalue corresponding to row i is r_i.

Thus, the elements of $D(\gamma)$ are defined by

$$d_i(\gamma) = (1-\gamma) + \gamma\left[2 + 2\cos\left(\frac{2\pi(T-i+1)}{2T+1}\right)\right]^{-1}. \tag{1.20}$$

No $T \times T$ matrix need be inverted, and estimation costs are comparable those of ordinary least squares with auto-regressive correction.

2. LARGE SAMPLE ANALYSIS

The intercept cannot be estimated consistently because it is continually subject to random changes. But, if γ where known, the maximum likelihood estimator of β would be efficient in the sense that it would have minimum variance in the class of unbiased estimators. In this section we prove that g, the m.l.e. of γ, is consistent implying $B(g)$, the m.l.e. of β, is asymptotically efficient. Further, the asymptotic distribution of $B(g)$ is normal with mean β and covariance $\sigma^2(X^1 Q_\gamma^{-1} X)^{-1}$.

Subsequently (γ_0, σ_0^2) will denote the true value of (γ, σ^2) and the T subscript will be implied for those elements depending upon the sample size. Letting S be the generalized sum of squared residuals condition on γ, the concentrated log likelihood function, (1.15), divided by $T/2$ is

$$L(\gamma; T) = -\ln S/T - T^{-1} \ln |D_\gamma| . \tag{2.1}$$

LEMMA A. *For* $\gamma \in [0, 1]$

$$\operatorname{plim} L(\gamma; T) = -\ln [\sigma_0^2 \textstyle\sum d_t(\gamma_0)/d_t(\gamma)] - T^{-1} \ln |D_\gamma| \equiv f(\gamma; T) \tag{2.2}$$

where here and subsequently summations are from 1 *to* T.

PROOF. Assuming the variables have been transformed via transformation P, the generalized sum of squared residuals is

$$S = w'[I - M_\gamma]w \tag{2.3}$$

where $w \sim N(0, \sigma_0^2 D_\gamma^{-1} D_0)$ and $M_\gamma = D_\gamma^{-1/2} X(X' D_\gamma^{-1} X)^{-1} X' D_\gamma^{-1/2}$. But from (1.20)

$$0 < d_t(\gamma_0)/d_t(\gamma) \leq 4 + \gamma_0/\gamma \tag{2.4}$$

which uniformly bounds the variances and fourth moments of w_t for $\gamma > 0$. This along with the fact M_γ is idempotent of rank k implies $w' M_\gamma w/T$ converges in mean to 0 and $w'w/T$ *to* $\sigma_0^2 \sum d_t(\gamma_0)/d_t(\gamma)$. As convergence in mean implies convergence in probability the result is proven for $\gamma > 0$.

If $\gamma = 0$ and $\gamma_0 > 0$

$$L(\gamma; T) - f(\gamma; T) = -\ln [\textstyle\sum w_t^2/\sigma_0^2 \sum d_t(\gamma_0)] .$$

The variance of the term in brackets is

$$3\sigma_0^{-2} \textstyle\sum d_t(\gamma_0)^2/[\sum d_t(\gamma_0)^2] .$$

Using well-known properties of eigenvalues [5, (273)], the numerator is of order T^3 as it equals the sum of all the elements of Q_γ squared while the denominator is of order T^4 as it is the square of the sum of the diagonal elements of Q_γ. Thus, the variance is of order T^{-1} implying convergence of the random variable to its mean in probability. If $\gamma = 0$ and $\gamma_0 = 0$, the result is trivial. Thus,

$$\operatorname{plim} [L(\gamma; T) - f(\gamma; T)] = -\ln [\textstyle\sum E w_t^2/\sigma_0^2 \sum d_t(\gamma_0)] = 0 ,$$

completing the proof.

REMARK. Pointwise convergence in probability does not imply convergence uniform in γ in probability. Uniform convergence (that is convergence in the L_∞-norm) is needed to conclude $L(\gamma; T)$ and $f(\gamma, T)$ have the same maximum in the limit. Lemma B establishes a continuity condition for the $L(\gamma; T)$ and

$f(\gamma;T)$ which implies uniform convergence given pointwise convergence. The final step in the proof is to show that $\lim f(\gamma;T)$ is continuous and has a unique maxima at γ_0. Then one can conclude that g converges in probability to γ.

LEMMA B. *The convergence of $L(\gamma;T)$ to $f(\gamma;T)$ is uniform in probability; that is*

$$\operatorname{plim} \sup_{\gamma\in[0,1]} |L(\gamma;T) - f(\gamma;T)| = 0 . \tag{2.5}$$

PROOF. Letting $e(\gamma) = Py - PXB(\gamma)$, then *for* $\gamma_2 > \gamma_1$

$$\begin{aligned} S(\gamma_2) \leq e(\gamma_1)'D_2^{-1}e(\gamma_1) &\leq [1 + 3(\gamma_2 - \gamma_1)]e(\gamma_1)'D_1^{-1}e(\gamma_1) \\ &= [1 + 3(\gamma_2 - \gamma_1)]S(\gamma_1) \end{aligned}$$

because $[1 + 3(\gamma_2 - \gamma_1)]D_1^{-1} - D_2^{-1}$ is positive definite, a result following from (1.20). Using this result

$$[S(\gamma_2) - S(\gamma_1)]/S(\gamma_1) \leq 3(\gamma_2 - \gamma_1)$$

which implies

$$\frac{d}{d\gamma}[L(\gamma;T) - \ln |D_\gamma|] \geq -3 . \tag{2.6}$$

Similarly,

$$\frac{d}{d\gamma}[f(\gamma;T) - \ln |D_\gamma| \geq -3 . \tag{2.7}$$

If $|L(\gamma_i;T) - f(\gamma_i;T)| \leq \varepsilon/3$ for $i = 1, 2$, then given (2.6) and (2.7)

$$|L(\gamma;T) - f(\gamma;T)| \leq 3(\gamma_1 - \gamma_2) + 2\varepsilon/3 \tag{2.8}$$

for all $\gamma_1 \leq \gamma \leq \gamma_2$. By Lemma A for any $\delta, \varepsilon > 0$ and N, there is a T^* such that for all $T \geq T^*$

$$\Pr\{|f(i/N;T) - L(i/N;T)| < \varepsilon/3\} \geq 1 - \delta/(N+1) \tag{2.9}$$

for $i = 0, 1, \ldots, N$. This implies

$$\Pr\{\max_i |f(i/N;T) - L(i/N;T)| < \varepsilon/3\} \geq 1 - \delta .$$

But, by (2.8) and (2.9), for $T \geq T^*$

$$\Pr\{\sup_{\gamma\in[0,1]} |f(\gamma;T) - L(\gamma;T)| \geq 2/3\varepsilon + 3/(N+1)\} \leq 1 - \delta .$$

Selecting $N > 8/\varepsilon$, we find

$$\Pr\{\sup_{\gamma\in[0,1]} |f(\gamma;T) - L(\gamma;T)| \geq \varepsilon\} \leq 1 - \delta .$$

This proves the Lemma.

LEMMA C. *There is a $K(\gamma;T) \geq K > 0$ such that*

$$f'(\gamma; T) = K(\gamma; T)(\gamma_0 - \gamma) \; for \; \gamma \in (0, 1].$$

PROOF. Differentiating $f(\gamma; T)$ yields

$$f'(\gamma; T) = \frac{1}{T} \sum \frac{d_t'(\gamma) d_t(\gamma_0)}{d_t(\gamma)^2} \Big/ \frac{1}{T} \sum \frac{d_t(\gamma_0)}{d_t(\gamma)} - \frac{1}{T} \sum \frac{d_t'(\gamma)}{d_t(\gamma)}.$$

Observe

$$f'(\gamma; T) = \left[\frac{1}{T} \sum \frac{d_t(\gamma_0)}{d_t(\gamma)}\right]^{-1} \frac{1}{T^2} \sum_{i,j} \frac{d_i'[d_i(\gamma_0) d_j(\gamma) - d_j(\gamma_0) d_i(\gamma)]}{d_i(\gamma)^2 d_j(\gamma)}$$

$$= \left[\frac{1}{T} \sum \frac{d_t(\gamma_0)}{d_t(\gamma)}\right]^{-1} \frac{\gamma_0 - \gamma}{2T^2 \gamma} \sum_{i,j} \frac{[d_i(\gamma) - d_j(\gamma)]^2}{d_i(\gamma)^2 d_j(\gamma)^2}.$$

From (1.20) $d_t(\gamma) > 1/4$, which implies the leading term exceeds $1/4$. Thus, with some additional algebra,

$$f'(\gamma; T) = K_1(\gamma; T)\,[T^{-1} \sum d_t(\gamma)^{-2} - (T^{-1} \sum d_t(\gamma)^{-1})^2](\gamma_0 - \gamma)$$

with $K_1(\gamma; T) > 1/8$. Using the definition of $d_t(\gamma)$, a positive lower bound can be obtained for the average squared deviation of the $d_t(\gamma)^{-1}$ from their average which holds for all T sufficiently large.

Lemma C implies that the functions $f(\gamma; T)$ have a unique maximum of γ_0. Furthermore, if $|\gamma - \gamma_0| > \varepsilon$, then $|f(\gamma_0; T) - f(\gamma; T)| \geq \varepsilon/K$. By Lemma B the probability that $L(\gamma; T)$ will be arbitrarily close to $f(\gamma, T)$ for all values of γ approaches one as T goes to infinity. Thus, the γ which maximize $L(\gamma; T)$ converges in probability to γ_0. This discussion can be summarized by the following theorem:

THEOREM. *The maximum likelihood estimator g of γ converges in probability to γ_0, the true parameter value.*

Following the usual analysis, but using the log likelihood function concentrated on β, the asymptotic distribution of $\theta' = (\gamma, \sigma^2)$ can be derived (c.f. [3]). The basic result is that $\sqrt{T}(\theta - \theta_0)$ is asymptotically normal with mean 0 and covariance equal to the inverse of the information matrix,

$$I(\theta_0) = \frac{1}{2T} \begin{bmatrix} \sum \left(\dfrac{d_i'(\gamma_0)}{d_i(\gamma_0)}\right)^2 & \dfrac{1}{\sigma_0^2} \sum \dfrac{d_i'(\gamma_0)}{d_i(\gamma_0)} \\ \dfrac{1}{\sigma_0^2} \sum \dfrac{d_i'(\gamma_0)}{d_i(\gamma_0)} & \dfrac{T}{\sigma_0^4} \end{bmatrix}.$$

REMARK. It is of interest to note that the only use of normality in the consistency proof was to insure the existence of the fourth moment of the w_t. The normality assumption merely provided a convenient function to be maximized.

Tufts University, U.S.A.,
Carnegie-Mellon University, U.S.A.

REFERENCES

[1] BOX, G. E. P. AND G. C. TIAO, "Changes in Level of a Non-stationary Time Series," *Biometrika*, LII (April, 1965), 181-192.
[2] COOLEY, T. F. AND E. C. PRESCOTT, "Test of the Adaptive Regression Model," *Review of Economics and Statistics*, (forthcoming).
[3] DHRYMES, P., *Distributed Lags: Problems of Formulation and Estimation* (San Francisco: Holden-Day, 1971).
[4] FRIEDMAN, M., *A Theory of the Consumption Function* (Princeton: Princeton University Press, 1957).
[5] HADLEY, G., *Linear Algebra* (Reading, Massachusetts: Addison-Wesley, 1961).
[6] MUTH, J., "Optimal Properties of Exponentially Weighted Forecasts" *Journal of the American Statistical Society*, LV (June, 1960), 299-306.
[7] TIAO, G. C. AND M. M. ALI, "Analysis of Correlated Random Effects: Linear Models with Two Random Component," Biometrika, LVIII (April, 1971), 37-51.