NBER WORKING PAPER SERIES

RIDGE ESTIMATORS FOR DISTRIBUTED LAG MODELS

G. S. Maddala*

Working Paper No. 69

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

October 1974

*NBER Computer Research Center. Research supported in part by National Science Foundation Grant GJ-1154X3 to the National Bureau of Economic Research, Inc.

Abstract

The paper explains how the Almon polynominal lag specification can be made stochastic in two different ways - one suggested by Shiller and another following the lines of Lindley and Smith. It is shown that both the estimators can be considered as modified ridge estimators. The paper then compares these modified ridge estimators with the ridge estimator suggested by Hoerl and Kennard. It is shown that for the estimation of distributed lag models the ridge estimator suggested by Hoerl and Kennard is not useful but that the modified ridge estimators corresponding to the stochastic versions of the Almon lag are promising. The paper has two empirical illustrations.

Acknowledgment

I would like to thank Kenneth M. Gaver for many helpful discussions and computational assistance.

Contents

Tables

1. Introduction

Consider a distributed lag model

$$y_t = \sum_{i=0}^{p} \beta_i x_{t-i} + u_t \tag{1}$$

where u_t are $IN(0,\sigma_u^2)$ $\qquad t = 1,2 \ldots n.$

The problems in the direct least squares estimation of (1) are: firstly, p the length of the lag is not known and secondly, even if p is known, because of high multicollinearity between the x_t, ordinary least squares estimates usually are erratic. The problem of an unknown p is usually 'solved' by assuming an infinite lag distribution that 'dies out' after a certain point. The Koyck [8], Solow [13], and Jorgenson [7] models are examples of this. There are, on the other hand, procedures that depend on a known p (or on the assumption that p can be determined by a bit of experimentation) like the models suggested by Almon [1], Leamer [9], and Shiller [12]. The basic problem these formulations are intended to solve is that of multicollinearity. The present paper assumes a known length of the lag distribution. We will write equation (1) in matrix notation as $y = X\beta + u$.

The plan of the paper is as follows: In section 2 we outline the Almon and Shiller procedures and show how they are related to the ridge estimators suggested by Hoerl and Kennard [5]. We also discuss a Bayesian version of the Almon estimator, different from that of Shiller, and show how it is related to the ridge estimators and the estimator suggested by Lindley and Smith [10]. The next section reports the results obtained by the application of these methods to two sets of data: one the set of data on capital appropriations and expenditures used by Almon and the other the data used by Griliches et al., [3] and Zellner and Geisel [16]. The final section presents the conclusions of the paper.

2. The Models Considered

(a) The Almon Method: Basically the procedure is based on the assumption that the β_i in (1) lie on a low degree polynomial. For the sake of illustration we will assume that it is a quadratic.

$$\beta_i = \alpha_o + \alpha_1 i + \alpha_2 i^2 \tag{2}$$

Then equation (1) can be written as

$$y_t = \sum_{i=0}^{p} (\alpha_o + \alpha_1 i + \alpha_2 i^2) x_{t-i} + u_t$$

$$= a_o z_{ot} + \alpha_1 z_{1t} + \alpha_2 z_{2t} + u_t$$

$$\text{where} \quad z_{jt} = \sum_{i=0}^{p} i^j x_{t-i} \tag{3}$$

One can also impose some end point constraints as Almon does, e.g., $\beta_i=0$ for i = -1 and p+1. This implies two linear restrictions on the α's in (2). It has been often argued that the nice smooth shape that is usually obtained with the Almon method is partly due to the imposition of the end point constraints.

Equation (2) can be written as

$$\beta = H\alpha \tag{4}$$

where
$$H = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ - & - & - \\ 1 & p & p^2 \end{bmatrix}$$

Define $M = I - H(H'H)^{-1}H'$ (5)

Then equation (4) implies the set of restrictions

$$M\beta = 0 \tag{6}$$

Thus, to get Almon's estimator we minimize $(y - X\beta)\ (y - X\beta)$ subject to (6). The resulting estimator is (Theil [15], p. 143)

$$\hat{\beta}_A = \hat{\beta} - (X'X)^{-1}\ M'\ [M(X'X)^{-1}\ M']^{-1}\ M\hat{\beta} \tag{7}$$

where $\hat{\beta}_A$ is the Almon estimator of β and $\hat{\beta}$ is the ordinary least squares estimator of β.

(b) Shiller's Method: The basic argument in Shiller's method is that we often specify a restriction such as (2) not because we believe in it but we believe the lag distribution to be smooth and consider (2) as an approximation.

One can add an error term to (2) and write

$$\beta_i = \alpha_o + \alpha_1 i + \alpha_2 i^2 + v_i \tag{8}$$

where v_i are $IN(0,\sigma_v^2)$. This will result in the model

$$y_t = \alpha_o z_{ot} + \alpha_1 z_{1t} + \alpha_2 z_{2t} + w_t$$

where z_{jt} are as defined in (3) and

$$w_t = u_t + \sum_{i=0}^{p} v_i x_{t-i}$$

This results in a complicated covariance matrix for the residuals. The residuals are heteroscedastic and autocorrelated. The Bayesian formulation of (8) will be discussed later.

Instead of making equation (2) stochastic, Shiller notes that assumption (2) implies that

$$\Delta^3 \beta_i = 0 \tag{9}$$

where $\Delta\beta_i = \beta_i - \beta_{i-1}$

Shiller makes equation (9) stochastic by adding an error term to it, i.e.,

$$\Delta^3 \beta_i = w_i \quad \text{where } w_i \text{ are } IN(0, \sigma_w^2) \tag{10}$$

This can be written as

$$R\beta = w \tag{11}$$

where

$$R = \begin{bmatrix} 1 & -3 & -3 & 1 & 0 & 0 & - & - & 0 \\ 0 & 1 & -3 & -3 & 1 & 0 & - & - & 0 \\ - & - & - & - & - & - & - & - & - \\ - & - & - & - & - & - & - & - & - \\ 0 & 0 & 0 & 0 & 0 & 1 & -3 & -3 & 1 \end{bmatrix}$$

R is a (p-1) × (p+1) matrix. Using the Theil-Goldberger [14] mixed estimation method, we get the estimator of β as

$$\hat{\beta}_s = [X'X + kR'R]^{-1} X'y \tag{12}$$

where $k = \frac{\sigma_u^2}{\sigma_w^2}$ is assumed known.

If we follow the non-stochastic version of Shiller's method, we minimize $(y - X\beta)\ (y - X\beta)$ subject to $R\beta = 0$ and the estimator of β we obtain is

$$\hat{\beta}_A = \hat{\beta} - (X'X)^{-1} R' [R(X'X)^{-1} R']^{-1} R \hat{\beta} \tag{13}$$

where $\hat{\beta}$ is the OLS estimator of β. This expression can be shown to be equivalent to the expression in (7). However when we introduce the stochastic term, it makes a difference whether we make (2) or (9) stochastic.

(c) Ridge Estimator: Heoerl and Kennard [5] present a set of biased least squares estimators which they call 'ridge estimators'. Their suggestion is to use, instead of the least squares estimator

$$\hat{\beta} = (X'X)^{-1} X'y$$

the modified estimator

$$\hat{\beta}_R = (X'X + kI) X'y \tag{14}$$

They suggest an iterative procedure to decide on a suitable k.

Instead of (14) we can consider a more general form

$$\hat{\beta}_R = (X'X + kQ)^{-1} X'y \tag{15}$$

where Q is a positive semi-definite matrix and written this way it can be seen that Shiller's estimator (12) is also a ridge estimator.

One can give a Bayesian interpretation to all these estimators and in fact this way of looking at the estimators is more revealing. Consider the ridge estimator (14). If the prior distribution of β is $N(0,\tau^2 I)$ then the mean of the posterior distribution of β is given by (14) with $k = \sigma^2/\tau^2$. Instead, if we assume the prior distribution of β to be $N(\delta,\tau^2\Delta)$ then the posterior distribution of β is also normal with mean

$$[X'X + k\Delta^{-1}]^{-1} [X'y + k\Delta^{-1}\delta] \tag{16}$$

and variance $\sigma^2(X'X + k\Delta^{-1})^{-1}$

If $\Delta = I$ then the posterior mean is

$$(X'X + kI)^{-1} (X'y + k\delta) \tag{17}$$

The important point to note is that the commonly used ridge estimator given by (14) implies a prior distribution for β with mean zero. This may not be a plausible assumption to make in many applications and if so equation (16) or (17) should be used. These modifications can be made very easily.

The Bayesian approach to the stochastic version of the Almon lag given by (8) can easily be seen to yield the ridge estimator as the posterior mean. Equation (8) can be written as

$$\beta = H\alpha + v$$

where H is defined in (4).

Following Lindley and Smith [10] we can say that

$$y \sim N(X\beta,\ I\,\sigma_u^2)$$

and $$\beta \sim N(H\alpha,\ I\,\sigma_v^2) \tag{18}$$

This still leaves us with the specification of the priors for α. We can assume a diffuse prior for α. Lindley and Smith prove the following theorem:

If $y \sim N(A_1\Theta_1,\ C_1)$

$\Theta_1 \sim N(A_2\Theta_2,\ C_2)$

and we assume a diffuse prior for Θ_2, then the posterior distribution of Θ_1 is $N(D_o d_o,\ D_o)$ where

$$D_o^{-1} = A_1'\,C_1^{-1}\,A_1 + C_2^{-1} - C_2^{-1}\,A_2(A_2'\,C_2^{-1}\,A_2)^{-1}\,A_2'\,C_2^{-1}$$

and $$d_o = A_1'\,C_1^{-1}\,y$$

In our case the posterior mean of β is (after simplification)

$$\beta^* = (X'X + kM)^{-1}\,X'y \tag{19}$$

where M is as defined in (5) and $k = \sigma_u^2/\sigma_v^2$.

Thus we see that the mean of the posterior distribution for the Bayesian version of the stochastic Almon lag specification is a ridge estimator of β.

As $\sigma_v^2 \to 0$ this should give the usual Almon estimator defined in (7). This can be checked as follows.

$$\beta^* = (X'X + kM)^{-1} X'y$$

$$= (X'X + kM^2)^{-1} X'y \quad \text{since M is idempotent.}$$

Now $(X'X + kM^2)^{-1}$ can be written as (see Rao [11] p.   )

$$(X'X)^{-1} - (X'X)^{-1} M [M'(X'X)^{-1} M + \frac{1}{k} I]^{-1} M' (X'X)^{-1}$$

As $\sigma_v^2 \to 0$, $\frac{1}{k} \to 0$ and thus β^* = the expression in (7).

Henceforth we will call the estimator (18) the Bayesian Almon estimator. The fact that $k = \sigma_u^2/\sigma_v^2$ suggests an iterative procedure for estimating k. We first estimate σ_u^2 from the least squares residuals and σ_v^2 from the estimated least square β's. In our computation of the Bayesian Almon estimator we used the iterative procedure suggested by Lindley and Smith ([10], p. 17).[1] In principle the same iterative procedure can also be used for Shiller's method.

(d) Lindley-Smith Estimator: Instead of making the assumption (18) Lindley and Smith assume

$$\beta \sim N(u1, \sigma_v^2 I)$$

where 1 is the unit vector and we have a vague prior for u. This is just a special case of the Almon lag (with zero degree polynomial). Thus, the posterior

---

[1]Throughout our discussion of the Bayesian Almon estimator, we assumed σ_u^2 and σ_v^2 known. When these are not known one has to assume priors for them. Since the analysis is similar to that given in Lindley and Smith we will not repeat it here.

mean they get is the same as in (19) except that instead of H we have the unit vector 1 and hence $M = I_{p+1} - \frac{1}{p+1} J_{p+1}$ when J is a matrix with all elements unity. This is the case considered by Lindley and Smith and for the sake of comparison we computed this estimator too using the iterative procedure suggested by them.

3. The Results

For illustrative purposes we used two sets of data: one the data used by Almon [1] and the other the data used by Griliches et al., [3]. The former data consist of 60 observations and the latter of 56 observations. Tables (1) and (2) present the correlation matrices for the two sets of data. These indicate how highly inter-correlated the variables are.

Tables (3) and (4) present the estimates obtained by using the OLS method, the Lindley-Smith method, the Bayes-Almon method and Shiller's method. In all cases the lag distribution was arbitrarily terminated at x_{t-8}. For the Bayes-Almon method we used a quadratic polynomial. As mentioned earlier, the Lindley-Smith method is a particular case of the Bayes-Almon method and corresponds to a zero-degree polynomial. Clearly the results given by this method are not as satisfactory as those obtained by a second degree polynomial. We report in the table the final value k* of k arrived at by the iterative procedure

and we also report the number of iterations after which things "converged." With the consumption function data the value of k did not increase much but with the Almon data the value of k kept on increasing. In both cases we terminated the iterations when the sum of the absolute values of the changes in the coefficients was less than .001.

For Shiller's method k was selected using the rule of thumb described in his paper viz. in $k = \frac{\sigma_u^2}{\sigma_v^2}$ we take σ_u^2 as the estimate of the residual variance from the OLS regression and σ_v^2 as $\frac{64S^2}{p^4}$ where S is the sum of the lag coefficients.

In our case p = 8 and $S \simeq 1$ so that $\sigma_v^2 \simeq \frac{1}{64}$. There is of course one problem with assuming a constant value of σ_v^2. In this case k will change with the units of measurement of y, since σ_u^2 changes. As mentioned earlier, the same iterative procedure used by Lindley and Smith can also be used for Shiller's method but we did not pursue this avenue yet. In our computations we used a first degree smoothness prior for the estimation by Shiller's method. This implies a first degree Almon polynomial.

We used the first degree smoothness prior because Shiller got good results with it in his paper and further it did capture the shape of a lag distribution that first rises and then declines. Since the results in Tables 3 and 4 do not enable us to make an adequate comparison between Shiller's method and the Bayes-Almon method, we recomputed the latter for the first degree polynomial. For the Almon data the coefficients were respectively: .12116, .12857, .14052, .13647, .12201, .10191, .09336, .07708, .06515. Sum = .98623 (convergence in 6 iterations). For the consumption function data the coefficients were respectively: .55104, .31501, .16258, -.15350, .00945, -.09091, .02169, .03385, .08655. Sum = .93576 (convergence in 7 iterations). These results show that there are substantial differences in the results produced by the two different ways of making the Almon polynomial specification stochastic. The Shiller procedure has produced a smoother lag distribution (particularly for the consumption function data) than the Bayes-Almon procedure but the estimates of the lag coefficients (particularly the initial ones) make more sense for the latter procedure. It is conceivable that the estimates produced by the two methods come closer if we used a similar iterative procedure for the Shiller method. But we have not pursued this further. It would be illuminating if we started with a known lag distribution and we are doing a Monte-Carlo study of the performance of the iterated and non-iterated estimators for the different specifications of the models discussed here. Hopefully this will shed more light on their relative performance.

The reason why the two pieces of data discussed here were selected is that there do exist estimates obtained by other methods (Koyck, Solow, etc.) for these data. These methods can be characterized as "strong parametric specifications" because they imply a strong specification that the β's lie on a particular shaped distribution. By contrast the methods discussed here are weak parametric specifications. We need not make an elaborate comparison here but the results obtained suggest that the strong parametric specifications may be responsible for producing some "plausible" - but distorted - lag shapes.

Finally, we come to the straight ridge method - the method of Hoerl and Kennard. We take

$$\hat{\beta}_R = (X'X + kI)^{-1} X'y$$

The usual procedure is to take X'X as the correlation matrix rather than the matrix of variances and covariances (to avoid problems with units of measurement). This merely amounts to multiplying the diagonal elements of the variance covariance matrix X'X by (1+k). We computed the ridge estimates by this method. We did not use an iterative procedure. Instead we used some trial values and one of the values of k suggested by Paul Holland [6] viz. k'_{a1}. We did not use the other k's suggested by him because they involve an iterative computation. Holland's suggestion was in the context of Robust regression but we simplified it for the simple ridge method here. However, his suggested k did not work too well. For the Almon data k'_{a1} turned out to be negative (-.031). For the consumption function data it was .026. But for this value of k, the ridge estimator really smoothed things out. The estimates of the parameters were: .11099, .10967, .10724, .10357, .10274, .10214, .10430, .10732, .11021. Sum = .95818. Apart from the fact that the sum is a bit high, the total is almost equally distributed among all the coefficients. Hence we decided to see what the results look like if k was reduced substantially. Table 5 presents the results for the consumption function data, for different values of k. Since our experience with the ridge regression is similar for the Almon data, we are not presenting the results here. The one puzzling feature in the results is the large and stable coefficient for x_{t-8}. This is perhaps a consequence of some seasonal elements in the data that we have not accounted for (a similar result did not appear for the Almon data). But apart from this, what the results in Table 5 (and similar results with Almon data not reported here) suggest is that the k for the simple Hoerl and Kennard ridge method has to be really very low. For values of k even as low as 0.1 or .005, the method really smooths things out. This also suggests the other modifications of the ridge method suggested earlier in the paper are more promising than the Hoerl and Kennard method.

4. Conclusions

The paper explains how two different stochastic formulations of the Almon polynomial method result in ridge estimators and how the Lindley-Smith procedure is also a special case of the Bayes-Almon method. All these methods have been applied for illustrative purposes to two sets of data. The results show that the ridge estimator of Hoerl and Kennard and the extension given by Lindley and Smith are not as promising for distributed lag estimation as the more general methods such as the stochastic versions of the Almon polynomial method. There are two ways of making the Almon polynomial method stochastic - one given by Shiller and the other which is a straightforward application of the Lindley and Smith procedure. The results presented, though inconclusive, suggest that the latter procedure is perhaps more flexible than Shiller's. Shiller argues that he does not have to specify a prior distribution for β - all he has to do is to specify a prior for differences in the β_i. However, this is only superficially true. If we assume a first degree smoothness prior for β this leaves two of the

β's (say β_o and β_1) free and implicitly Shiller is assuming a diffuse prior for these parameters. Anyway the relative merits of the two procedures and the iterative versus non-iterative computation of k need more detailed study and the results of some further empirical examples and some Monte Carlo experiments which are under way will be presented elsewhere.

Table 1: Correlation Matrix: Almon Data

1	1.000									
2	0.962	1.000								
3	0.910	0.961	1.000							
4	0.845	0.908	0.960	1.000						
5	0.769	0.842	0.906	0.959	1.000					
6	0.700	0.763	0.837	0.903	0.958	1.000				
7	0.630	0.689	0.752	0.829	0.899	0.955	1.000			
8	0.583	0.619	0.680	0.748	0.832	0.899	0.957	1.000		
9	0.521	0.569	0.607	0.671	0.743	0.824	0.891	0.950	1.000	
10	0.866	0.919	0.958	0.974	0.965	0.934	0.887	0.829	0.759	1.000

Table 2: Correlation Matrix: Consumption Function Data

1	1.000									
2	0.995	1.000								
3	0.991	0.995	1.000							
4	0.986	0.991	0.995	1.000						
5	0.982	0.986	0.990	0.995	1.000					
6	0.980	0.982	0.985	0.990	0.995	1.000				
7	0.980	0.980	0.981	0.985	0.990	0.995	1.000			
8	0.982	0.980	0.980	0.981	0.985	0.990	0.995	1.000		
9	0.981	0.981	0.980	0.979	0.980	0.985	0.990	0.994	1.000	
10	0.994	0.991	0.986	0.978	0.975	0.973	0.975	0.978	0.979	1.000

Note: Variable 10 is the dependent variable y_t. Variables 1 to 9 are x_t to x_{t-8}.

Table 3: Estimates for Almon Data

Lag	OLS	Lindley-Smith $k^* = 193.5$ (After 8 iter.)	Bayes-Almon $k^* = 12.35$ (After 5 iter.)	Shiller $k = 729.326$
0	.07272	.11441	.09142	.13251
1	.08121	.11648	.12446	.13209
2	.23184	.11845	.15607	.13052
3	.18436	.11698	.15540	.12631
4	.13406	.11314	.13508	.11868
5	.01382	.10820	.10678	.10778
6	.13647	.10517	.09980	.09445
7	.06380	.10220	.07331	.07962
8	.06870	.10086	.04225	.06423
Sum	.98698	.99589	.98457	.98619

Table 4: Estimates for the Consumption Function Data

Lag	OLS	Lindley-Smith k* = .0364 (After 9 iter.)	Bayes-Almon k* = .0182 (After 5 iter.)	Shiller k = 329.466
0	.70974	.47681	.64012	.28324
1	.20808	.29430	.29529	.23784
2	.27463	.15736	.14319	.19255
3	-.48068	-.10832	-.21783	.14755
4	.25129	.00044	.02372	.10296
5	-.23845	.08088	-.12591	.05883
6	.12432	.01800	.00688	.01508
7	-.11278	.05833	.00925	-.02843
8	.19838	.12451	.16081	-.07183
Sum	.93453	1.10231	.93552	.93780

Table 5: Ridge Estimates for Consumption Function Data

	Value of k					
Lag	0.0	.0002	.0006	.0010	.0014	.0020
0	.70974	.42246	.29302	.24038	.21096	.18489
1	.20808	.28187	.22554	.19578	.17773	.16096
2	.27463	.15615	.14612	.13865	.13324	.12764
3	-.48068	-.06079	.03052	.05761	.07060	.08088
4	.25129	-.00301	.02429	.04473	.05736	.06902
5	-.23845	-.06461	-.00562	.02304	.04010	.05578
6	.12432	.01705	.03600	.05116	.06135	.07138
7	-.11278	.06733	.07964	.08491	.08862	.09254
8	.19838	.12632	.11941	.11563	.11367	.11220
Sum	.93453	.94277	.94892	.95189	.95363	.95529

References

[1] Almon, S. - "The Distributed Lag Between Capital Appropriations and Expenditures" - Econometrica, 1965, pp. 178-196.

[2] Dhrymes, P. J. - Distributed Lags: Problems of Estimation and Formulation (Holden Day, San Francisco), 1971.

[3] Griliches, Z., G. S. Maddala, R. E. Lucas and N. Wallace - "Notes on Estimated Aggregate Quarterly Consumption Functions" - Econometrica, 1962, pp. 491-500.

[4] Griliches, Z. - "Distributed Lags: A Survey" - Econometrica, 1967, pp. 16-49.

[5] Hoerl, A. E., and R. W. Kennard - "Ridge Regression: Biased Estimation for Non-orthogonal Problems" - Technometrics, 1970, pp. 55-67.

[6] Holland, Paul W. - "Weighted Ridge Regression - Combining Ridge and Robust Regression Methods" - Working Paper No. 11, National Bureau of Economic Research, September 1973.

[7] Jorgenson, D. W. - "Rational Distributed Lag Functions" - Econometrica, 1966, pp. 135-149.

[8] Koyck, L. M. - Distributed Lags and Investment Analysis, (North Holland Publishing Company, Amsterdam, 1954).

[9] Leamer, Edward E. - "A Class of Informative Priors and Distributed Lag Analysis" - Econometrica, 1972, pp. 1059-1081.

[10] Lindley, D. V. and A. F. M. Smith - "Bayes Estimates for the Linear Model" (with Discussion) - Journal of the Royal Statistical Society, B Series, 1972, pp. 1-41.

[11] Rao, C. R. - Linear Statistical Inference and Its Applications, (Wiley, New York), 1965.

[12] Shiller, Robert J. - "A Distributed Lag Estimator Derived from Smoothness Priors" - Econometrica, 1973, pp. 775-788.

[13] Solow, Robert M. - "On a Family of Lag Distributions" - Econometrica, 1960, pp. 393-406.

[14] Theil, H. and A. S. Goldberger - "On Pure and Mixed Statistical Estimation in Economics" - International Economic Review, 1960, pp. 65-78.

[15] Theil, H. - Principles of Econometrics (Wiley, New York), 1971.

[16] Zellner, A. and M. S. Geisel - "Analysis of Distributed Lag Models With Applications to Consumption Function Estimation" - Econometrica, 1970, pp. 865-888.