Continental J. Applied Sciences 8 (2): 26 - 33, 2013 ISSN: 1597 – 9928
© Wilolud Journals, 2013 http://www.wiloludjournal.com
Printed in Nigeria doi:10.5707/cjapplsci.2013.8.2.26.33

LISTING OF ALL SUBSETS AND SELECTION OF BEST SUBSET IN A FINITE DISTRIBUTED LAG MODEL USING METEOROLOGICAL DATA

Ojo, J. F. and Femi J. Ayoola
Department of Statistics, University of Ibadan, Ibadan, Nigeria.

ABSTRACT
This paper develops algorithms for listing and estimation of all subsets in a finite distributed lag model, with a view to select the best subset from the listed subsets thereby removing the parameters that contribute insignificantly to the model and more so reduce the problem of multi-collinearity. We describe the method of estimation for full and subset finite distributed lag models. In order to select the best order for full and subset finite distributed lag models, Akaike information criterion (AIC) is adopted. The estimation technique is illustrated with respect to a meteorological series where rainfall is the dependent variable and temperature is the explanatory variable. The best order for the full finite distributed lag model is achieved at lag six (6) and the 2^q-1 for all subsets gave a total of sixty three (63) subsets that is 2^6-1. The best order having estimated the sixty three equations occurred at the 31st subsets. This led to the removal of insignificant parameters in the subset model. The subset model is more appropriate as the value for the AIC in the subset model is less than the AIC value in the full model. Apart from the intercept, the full finite distributed lag model had six parameters while the subset model had three parameters meaning that those parameters that contributed insignificantly to the model have been removed thereby leading to a better model and a reduction in the problem of multi-collinearity.

KEYWORDS:Full and subset finite distributed lag models, Multi-collinearity, Parameters, Akaike information criterion, Meteorological series.

Received for Publication: 13/02/13 Acceptedfor Publication: 10/04/13
Corresponding Author: jfunminiyiojo@yahoo.co.uk

INTRODUCTION

Economic decisions have consequences that may last a long time. When the income tax is increased, consumers have less disposable income, reducing their expenditures on goods and services, which reduces profits of suppliers, which reduces the demand for productive inputs, which reduces the profits of the input suppliers. These effects do not occur instantaneously but are spread, or distributed, over future time periods. Economic actions or decisions taken at one point in time, t, have effects on the economy at time t, but also at times t+1, t+2, and so on (Judge et al. 2000). Algebraically, we can represent this lag effect by saying that a change in a policy variable x_t has an effect upon economic outcomes $y_t, y_{t+1}, y_{t+2}, \ldots\ldots..$ If we turn this around slightly, then we can say that y_t is affected by the values of $x_t, x_{t-1}, x_{t-2}, \ldots\ldots.$, or

$$y_t = f(x_t, x_{t-1}, x_{t-2}, \ldots\ldots) \qquad (1)$$

Models like (1) are said to be dynamicsince they describe the evolving economy and its reactions over time. One immediate question with models like (1) is how far back in time we must go, or the length of the distributed lag. Infinite distributed lag models portray the effects as lasting, essentially, forever. In finite distributed lagmodels we assume that the effect of a change in a (policy) variable x_t affects economic outcomes y_t only for a certain, fixed, period of time. This distributed lag model is finite as the duration of the effects is a finite period of time, namely n periods (Judge et al. 2000).

There may be more parameters in the full finite distributed lag models and some of these parameters may contribute insignificantly to the model. Therefore looking for subset finite distributed lag models where these parameters are removed is highly desirable and it is expected that inference from the subset model is always better that the inference from the full model. The theory of subsetting has been considered by Hocking and Leslie (1967) where selection of the best subset in regression analysis was treated. Also Furnival (1971), all possible regressions with less computation and Mallows (1973), treated choosing subset regression.

When the lag length (n) is long in finite distributed lag model, severe multi-collinearity can occur (Judge et al. 2000). With the possibility of more parameters in the full finite distributed lag model, it is desirable to look for subset finite distributed lag model which will have fewer parameters. Listing all subsets in the long lag length (n) of full finite distributed lag model and selecting the best subset model using selection criterion from these subsets will reduce the lag length drastically thereby lessening the seriousness of multi-collinearity. A selection criterion is used to specify the lag length (full and subset model). One of the most popular criteria is the Akaike Information Criterion (AIC) in Akaike (1974) which we shall adopt in this study

In a nutshell, subset model with fewer parameters is desirable for a finite distributed lag model so as to reduce the problem of multi-collinearity which always characterized a finite distributed lag model when the lag length is long. This study is set to fill this gap using subsetting approach.

MATERIALS AND METHODS

Finite Distributed Lag Models

Instatisticsandeconometrics, a distributed lag model is a model fortime series data in which a regression equation is used to predict current values of adependent variable based on both the current values of an explanatory variable and the lagged (past period) values of this explanatory variable in Cromwell et al. (1994) and Judge et al.(1980)

In order to convert (1) into an econometric model we must choose a functional form, add an error term and make assumptions about the properties of the error term. As a first approximation let us assume that the functional form is linear, so that the finite distributed lag model, with an additive error term, is

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \dots\dots + \beta_n x_{t-n} + e_t \qquad (2)$$

where we assume that $E(e_t) = 0,\ \mathrm{var}(e_t) = \sigma^2,\ \mathrm{cov}(e_t, e_s) = 0$. Note that if we have T observations on the pairs (y_t, x_t) then only $T - n$ complete observations are available for estimation since n observations are "lost" in creating $x_{t-1}, x_{t-2}, \dots, x_{t-n}$. In this finite distributed lag the parameter α is the intercept and the parameter β_i is called a distributed lag weight. When $\beta_i$are known, estimating α is very easy.

Estimation of a Finite Distributed Lag Model

Theorem 1: Equation 2 can be written in a notation form as $y = X\beta + e$. Let $y = X\beta + e$, where X: n x p is of full rank, β: p x 1 is a vector of unknown parameters, and e : n x 1 is a random vector with mean 0 and $\text{cov}(e, e') = \text{var}(e) = \sigma^2 I$. The least squares estimator for β is denoted by b and is given by

$$b = (X'X)^{-1} X'y \qquad (3)$$

Proof: The vector of residuals can be written as $\hat{e} = y - Xb$ and hence

$$\hat{e}'\hat{e} = \sum_{i=1}^{n} \hat{e}_i^2 = (y - Xb)'(y - Xb)$$
$$= y'y - y'Xb - b'X'y + b'X'Xb$$

Since $b'X'y$ is 1 x 1, $(b'X'y)' = y'Xb$.

Hence $\hat{e}'\hat{e} = y'y - 2y'Xb + b'X'Xb$.

Applying the rules for differentiation we have

$$\frac{\partial}{\partial b}\hat{e}'\hat{e} = -2X'y + 2(X'Xb).$$

Setting this derivatives equal to zero, we obtain

$$-2X'y + 2(X'X)b = 0$$

or

$$(X'X)b = X'y.$$

This expression is called the Normal equation. Hence

$$b = (X'X)^{-1} X'y.$$

Let

$$SSE = \hat{e}'\hat{e} = (y - Xb)'(y - Xb).$$

Now

$$y - Xb = y - X(X'X)^{-1} X'y = \left(I - X(X'X)^{-1} X'\right)y = (I - P)y,$$

where $P = X(X'X)^{-1} X'$

Hence

$$SSE = \left((I - P)y\right)'(I - P)y = y'(I - P)y \qquad (4)$$

Selection of the Length of the Finite Lag
Numerous procedures have been suggested for selecting the length *n* of a finite distributed lag in Judge et al. (2000). Two goodness-of-fit measures that are more appropriate are Akaike's*AIC* criterion

$$AIC = \ln \frac{SSE_n}{T-N} + \frac{2(n+2)}{T-N} \qquad (5)$$

Schwarz's *SC* criterion

$$SC(n) = \ln \frac{SSE_n}{T-N} + \frac{(n+2)\ln(T-N)}{T-N} \qquad (6)$$

For each of these measures we seek that lag length n^* that minimizes the criterion used. Since adding more lagged variables reduces *SSE*, the second part of each of the criteria is a penalty function for adding additional lags. These measures weigh reductions in sum of squared errors obtained by adding additional lags against the penalty imposed by each. They are useful for comparing lag lengths of alternative models estimated using the same number of observations. In this study we shall use AIC criterion for selecting best order in full and subset models.

Algorithm for 2^q-1 Possible Subsets in Finite Distributed Lag Models
The 2^q-1subsets of a finite distributed lag models make use of the properties of permutation and combinatorial analysis and the algorithm goes thus:

The q value, which is an integer and a maximum lag, is selected from the fitted models using a selection criterion. The first sets of numbers are one digit number 1, 2, 3, up to q. The second sets of numbers are 2-digit numbers arranged in such a way that the first digit 1 is taken and combined with the next number until q is reached; the next digit 2 is picked and combines with the next number until q is reached, the next digit with the next number until q is reached. This is continued until (q-1) and q are reached.

The third sets of numbers are three digits and the second digit is operated on, to produce third digit. Our guide is that the 2-digit must have the next forward number until q is reached and the next two digits must have the next forward digit until q is reached. This is continued until we have (q-2)(q-1) and q. This is continued until we have the maximum digit, which is the digit q that is (q-n)(q-s)(q-u)(q-z)(q-m)(q-y),…………,q. where n<q by 1, s<q by 2, u<q by 3,…………………..,q. Suppose we have our q to be 4, (the best order using selection criterion) there are $2^4 - 1$ possible subsets that is 15 subsets. Following our algorithm we shall have the following: 1; 2; 3; 4; (1,2); (1,3); (1,4); (2,3); (2,4); (3,4); (1,2,3); (1,2,4); (1,3,4); (2,3,4); (1,2,3,4). The subset equations are

$$y_t = \alpha + \beta_0 x_t + e_t \qquad (7)$$

$$y_t = \alpha + \beta_1 x_{t-1} + e_t \qquad (8)$$

$$y_t = \alpha + \beta_2 x_{t-2} + e_t \qquad (9)$$

$$y_t = \alpha + \beta_3 x_{t-3} + e_t \qquad (10)$$

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + e_t \quad (11)$$

$$y_t = \alpha + \beta_0 x_t + \beta_2 x_{t-2} + e_t \quad (12)$$

$$y_t = \alpha + \beta_0 x_t + \beta_3 x_{t-3} + e_t \quad (13)$$

$$y_t = \alpha + \beta_1 x_{t-1} + \beta_2 x_{t-2} + e_t \quad (14)$$

$$y_t = \alpha + \beta_1 x_{t-1} + \beta_3 x_{t-3} + e_t \quad (15)$$

$$y_t = \alpha + \beta_2 x_{t-2} + \beta_3 x_{t-3} + e_t \quad (16)$$

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + e_t \quad (17)$$

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_3 x_{t-3} + e_t \quad (18)$$

$$y_t = \alpha + \beta_0 x_t + \beta_2 x_{t-2} + \beta_3 x_{t-3} + e_t \quad (19)$$

$$y_t = \alpha + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + e_t \quad (20)$$

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + e_t \quad (21)$$

where $1=\beta_0$ $2=\beta_1$, $3=\beta_2$, $4=\beta_3$. If the minimum AIC occurred in (15) that model is the best subset finite distributed lag model.

Description of Algorithm for Fitting Best Subset Finite Distributed Lag Model

For the sake of simplicity, we will break the algorithm down into the following steps.

Step 1

Fit various order of finite distributed lag model of the form using the estimation technique in section 2.2

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots\ldots + \beta_n x_{t-n} + e_t$$

Step 2

Choose the model for which Akaike Information Criterion (AIC) is minimum among various order fitted in step 1 and let the AIC value be AIC(1) and R^2 be $R^2(1)$.

Step 3

Fit possible subsets of chosen model in step 2 using $2^q - 1$ subsets approach described in section 2.3 and the estimation technique in section 2.2.

Step 4

Choose the model for which AIC is minimum among the fitted models in step 3 and let the AIC value be AIC(2) and R^2 be $R^2(2)$. When $AIC(2) < (AIC(1)$ and $R^2(2) < R^2(1)$ then we have the best subset model and at this point the coefficients that do not contribute significantly to the model are removed.

RESULTS AND DISCUSSION

Scope of the Study

The data for this study was secondary obtained from Institute of Agricultural Research and Technology, (IART), Ibadan and Forestry Research Institute of Nigeria (FRIN), Ibadan, Nigeria between 1979 and 2008. Rainfall series is the dependent variable while temperature series is the explanatory variable.

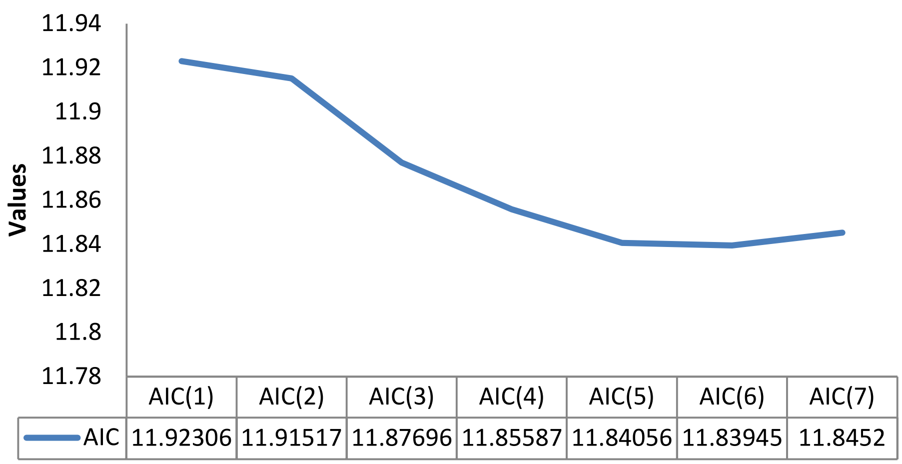
Fitting of Full Finite Distributed Lag Models

The full finite distributed lag model of order one to seven are fitted. The choice of the order is made on the basis of the Akaike information criterion. It is found that AIC is minimum at order six. The value of the AIC are recorded as follows:

Table 1: AIC Value for Full Model

Order	AIC
1	11.92306
2	11.91517
3	11.87696
4	11.85587
5	11.84056
6	11.83945
7	11.84520

Figure 1: AIC Values for Full Finite Distributed Lag Model



The estimated values of the coefficient of the model are as follows α=289.0294, β_0=-11.79314, β_1=-2.363999, β_2=2.272282, β_3=1.211662, β_4=2.461767, β_5=2.866697.

The model is given as follows:

$$\hat{y}_t = 289.0294 - 11.79314\, x_t - 2.363999\, x_{t-1} + 2.272282\, x_{t-2} + 1.211662\, x_{t-3} + 2.461767\, x_{t-4} + 2.866697\, x_{t-5}$$

Listing and Fitting of Subset Finite Distributed Lag Model

The method described in section 2.4 and 2.5 were employed in fitting the best subset finite distributed lag model. There are 2^6-1= 63 possible subsets which gave 63 equations. The 63 subsets were listed from the following notation, 1=β_0, 2=β_1, 3=β_2, 4=β_3, 5= β_4 6=β_5. These equations were estimated and the AIC values were recorded. AIC table and graph of the AIC values are shown below. The choice of the order is made on the basis of minimum AIC and having considered the AIC attached to the 63 equations; it was found that AIC is minimum at subsets 31 denoted as (1, 5, 6) that is β_0, β_4 and β_5 and the estimated model is given as $\hat{y}_t = 288.7018 - 12.14200x_t + 3.987625x_{t-4} + 2.819912x_{t-5}$

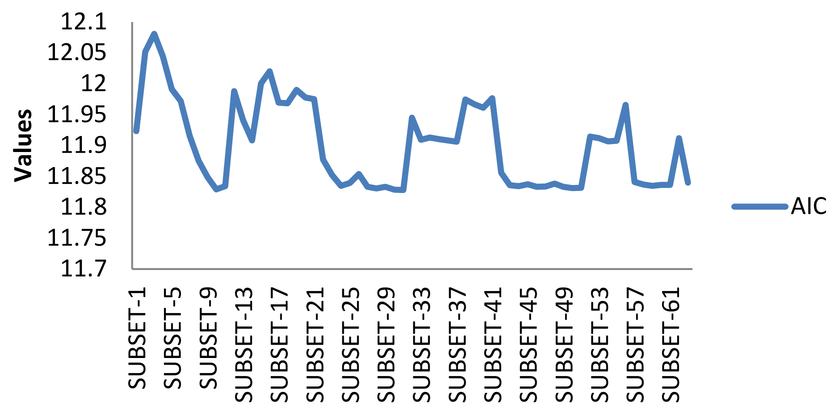
The minimum AIC value for the best subset finite distributed lag model is 11.82789 which is less than 11.83945 the value for the full finite distributed lag model. From the best subset finite distributed lag variable we could see that the coefficients that do not contribute significantly to the model have been removed using subsetting approach. From the AIC value best subset finite distributed lag model is more appropriate than full finite distributed lag model.

Apart from the intercept, the full finite distributed lag model had six parameters while the subset model had three parameters meaning that those parameters that contributed insignificantly to the model have been removed thereby leading to a better model and a reduction in the problem of multi-collinearity.

Table 2: AIC Values for all Subsets (2^6-1)

Subset	AIC	Subset	AIC	Subset	AIC
1	11.92306	22	11.87696	43	11.83543
2	12.05159	23	11.85213	44	11.83410
3	12.08071	24	11.83421	45	11.83688
4	12.04436	25	11.83905	46	11.83298
5	11.99068	26	11.85367	47	11.83355
6	11.97195	27	11.83280	48	11.83798
7	11.91517	28	11.83042	49	11.83290
8	11.87470	29	11.83262	50	11.83103
9	11.84821	30	11.82796	51	11.83112
10	11.82885	31	11.82789	52	11.91401
11	11.83414	32	11.94479	53	11.91169
12	11.98796	33	11.90928	54	11.90668
13	11.94077	34	11.91282	55	11.90768
14	11.90818	35	11.91009	56	11.96532
15	12.00001	36	11.90810	57	11.84056
16	12.02008	37	11.90604	58	11.83636
17	11.96930	38	11.97409	59	11.83436
18	11.96783	39	11.96638	60	11.83599
19	11.98968	40	11.96058	61	11.83617
20	11.97743	41	11.97644	62	11.91137
21	11.97507	42	11.85587	63	11.83945

Figure 2: Graph of AIC for All Subsets



CONCLUSION

This study has developed algorithms for listing and estimation of all subsets in a finite distributed lag model, with a view to select the best subset from the listed subsets thereby removing the parameters that contribute insignificantly to the model and more so reduce the problem of multi-collinearity. The parameters of full and subset finite distributed lag models were estimated. The estimation technique is illustrated with respect to a meteorological series where rainfall is the dependent variable and temperature is the explanatory variable. The best order for the full finite distributed lag model is achieved at lag six (6) while that of subset occurred at subset 31. The subset model is more appropriate as the value for the AIC in the subset is less than the AIC value in the full model. Apart from the intercept, the full finite distributed lag model had six parameters while the subset model had three parameters meaning that those parameters that contributed insignificantly to the model have been removed thereby leading to a better model and a reduction in the problem of multi-collinearity.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6), 716-723.

Cromwell J. B., Walter C. L., Michael J. H., Michel T. (1994). *Multivariate Tests for Time Series Models*. SAGE Publications.

Furnival, G. M. (1971). All possible Regressions with Less Computation.*Technometrics* 13, 403-408.

Hocking, R. R. and Leslie R. N. (1967).Selection of the Best Subset in Regression Analysis.*Technometrics* 9, 531-540

Judge G. G., William E. G., Carter R. H., Helmit L., Tsoung-Chao L. (1980). *The Theory and Practice of Econometrics*.Wiley Publication.

Judge G. G., Carter R. H., William E. G. (2000). *Undergraduate Econometrics*, 2nd Edition, John Wiley Publication.

Mallows, C. L. (1973).Choosing Subset Regression.*Technometrics*17, 213-219.