POLYNOMIAL DISTRIBUTED LAGS:
A UNIFIED TREATMENT

By PRAVIN K. TRIVEDI AND ADRIAN R. PAGAN

1. Introduction

The estimation of finite distributed lags by the technique of polynomial approximation developed by Almon [3] has become a standard feature of the econometric scene in recent years. Early on, it was realized that the extreme simplicity of the method masked some deep problems e.g., as Dhrymes [10] and Trivedi [30] discovered, the imposition of zero restrictions at the ends of the lag distribution could, if applied uncritically, yield very misleading results. However, after the cessation of these investigations, there seems to have been little attention paid to the pitfalls of the Almon technique until the recent articles by Schmidt and Waud [22] and Godfrey and Poskitt [13], although there does appear to have been an oral, if not a written tradition concerning the issues raised in these papers. In what follows, the substantive results from these articles are synthesized by utilizing the theory of linear restrictions, such a viewpoint leading naturally to a re-evaluation of the role of the technique in the estimation of lag distributions.

The problem is to estimate the finite distributed lag model

\[ y_t = \sum_{j=0}^{n} w_j x_{t-j} + u_t \quad t = 1, \ldots, T \]

where the coefficients \( w_j \) are assumed to lie on the \( p \)th order polynomial

\[ w_j = \alpha_0 + \alpha_1 j + \cdots + \alpha_p j^p. \]

and \( u \sim N(0, \sigma^2 I) \).

Equations (1) and (2) may be written in matrix form as

\[ y = Xw + u \]

\[ w = J\alpha \]

where \( X \) is a \( T \times (n+1) \) matrix containing \( x_{t-j} \) as columns, \( w' = [w_0 \ldots w_n], \alpha' = [\alpha_0 \ldots \alpha_p] \) and \( J \) is a \( (n+1) \times (p+1) \) Vandermonde matrix with \( (i-1)j^{i-1} \) in the \( i^{th} \) row, \( j^{th} \) column \((i = 1, \ldots, n+1; j = 1, \ldots, p+1)\).

Because \( w = J\alpha \) it is clear that there are \( (n-p) \) linear restrictions imposed upon \( w \) and Shiller \([25]\) has pointed out that these restrictions are the set formed by setting the \( (p+1) \)th differences of the parameters to zero i.e. \( (1 - z)^{p+1}w_j = 0 \) where \( z \) is a lag operator. Such linear restrictions may be written in the homogeneous form

\[ Rw = 0 \]

where \( R \) is a \( (n-p) \times (n+1) \) matrix whose \((i, i+j)^{th}\) element is the coefficient of \( z^j \) in the

---

1) The method outlined above is generally identified as the direct rather than Almon method of estimating lag distributions—see Cooper [8] for the distinction. In a later section this distinction becomes important, but merely for computational reasons.
expansion of \((1 - z)^{p+1}\) \((i = 1, \ldots, n - p; j = 0, \ldots, p + 1)\) with all remaining elements being zero.

The derivation of the restricted least squares estimate of \(w, \hat{w}\), is a simple matter discussed in most texts e.g. Theil [28]. Denoting the OLS (unrestricted) estimator by \(\hat{w} = (X'X)^{-1}X'y\),

\[(6) \quad \hat{w} = \hat{w} - MR\hat{w}\]

where
\[
M = (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}.
\]

2. The Effects of Mis-specification

Even if the strong assumption that the weights \(w\) lie exactly upon a polynomial of degree \(p\) were accepted as a reasonable starting point (and we will devote some time to this in a later section), it still remains to choose \(p\) and \(n\); hence the necessity to analyse the consequences of an incorrect choice arises. Four cases are examined in detail; two of which involve choosing \(n\) incorrectly given the correct choice of \(p\), while the other two center on an incorrect choice of \(p\) given a correct \(n\). Throughout, \(n^*\) and \(p^*\) denote the true values of \(n\) and \(p\), so that the four cases considered are

I: \(n > n^*, p = p^*\); II: \(n < n^*, p = p^*\); III: \(n = n^*, p > p^*\);

IV: \(n = n^*, p < p^*\).

I: \(n > n^*, p = p^*\)

**Proposition:** If the assumed lag length \(n\) overstates the true lag length \((n^*)\) by more than the degree of the approximating polynomial \((p = p^*)\), the polynomial distributed lag (PDL) estimator is biased.

**Proof:** Let \(n^* > p^*\). Then the new model is

\[(7) \quad y = X_L\hat{w}_L + u_L\]
\[(8) \quad \hat{R}_L\hat{w}_L = 0\]

where \(\hat{w}_L^* = [\hat{w}^*\hat{w}_\ast], \hat{w}\) is \((n + 1) \times 1, \hat{w}_\ast\) is \((n - n^*) \times 1\) and \(X_L\) contains all lags of \(x_t\) up to \(n\). We may partition \(\hat{R}_L\) conformably with \(\hat{w}_L\) as

\[
\hat{R}_L = \begin{pmatrix}
R & r_{12} \\
r_{21} & r_{22}
\end{pmatrix}.
\]

Then, if \(\hat{w}_L = (X_L'X_L)^{-1}X_L'y\), the PDL estimator is

\[(9) \quad \hat{w}_L = \hat{w}_L - \hat{M}_L\hat{R}_L\hat{w}_L\]

where

\[(10) \quad \hat{M}_L = (X_L'X_L)^{-1}\hat{R}_L'(\hat{R}_L (X_L'X_L)^{-1}\hat{R}_L')^{-1}.\]

Expanding (10) to obtain

\[
\hat{w}_L = \begin{pmatrix}
\hat{w} \\
\hat{w}_\ast
\end{pmatrix} - \hat{M}_L \begin{pmatrix}
R & r_{12} \\
r_{21} & r_{22}
\end{pmatrix} \begin{pmatrix}
\hat{w} \\
\hat{w}_\ast
\end{pmatrix}
\]

and taking expectations

\[2) \quad \text{Throughout the paper } \hat{M} \text{ with a subscript refers to an expression such as this, i.e., } \hat{M}_j = (X_j'X_j)^{-1}(R_j'(X_j'X_j)^{-1}R_j')^{-1}. \text{ Also } X \text{ is generally taken to be non-stochastic with } X'X \text{ having full rank.}\]
April 1979  P.K. Trivedi and A.R. Pagan: Polynomial Distributed Lags

\[
E(\tilde{w}_L) = \begin{pmatrix} w \\ \vdots \\ 0 \end{pmatrix} - M_L \begin{pmatrix} 0 \\ \vdots \\ r_{21} \end{pmatrix}
\]

since, under the null hypothesis \( Rw = 0 \) and \( w^* = 0 \),

(i) \( E(R\tilde{w}) = Rw = 0 \)

(ii) \( E(r_{12}\tilde{w}^*) = r_{12}w^* = 0 \)

(iii) \( E(r_{21}\tilde{w}) = r_{21}w = 0 \)

(iv) \( E(r_{22}\tilde{w}^*) = r_{22}w^* = 0 \)

The question of bias essentially revolves around the solution to the homogeneous equations

\[
Rw = 0 \quad (n^* - p^* \text{ equations})
\]

\[
r_{21}w = 0 \quad (n - n^* \text{ equations}),
\]

that is, whether the solution to the first \((n^* - p^*)\) homogeneous equations is in any way affected by the addition of \((n - n^*)\) incorrect restrictions. If it is, we shall say that estimation bias is introduced. Since the rank of \( R \) is \((n^* - p^*) < n \) a non-trivial solution to (i) exists. A necessary and sufficient condition for a non-trivial solution to (i) and (iii) jointly is that the rank of

\[
\begin{pmatrix} R \\ r_{21} \end{pmatrix} < n.
\]

This condition is violated when \((n - n^*) > p^* + 1\), so only a trivial solution exists and hence the estimates will be biassed. If \((n - n^*) \leq p\) then a non-trivial solution exists but we cannot be sure that it will be the same as the one obtained by solving (i) alone. Hence we cannot be certain that the estimator is unbiased. The proof is easily modified to take account of end-point restrictions. If \( n^* = p^* \), the \( R_L \) matrix must be partitioned column-wise, not row-wise, to demonstrate the proposition.

II: \( n < n^*, p = p^* \)

**Proposition:** Understating the true lag length generally leads to a bias in the PDL estimator.

If the mis-specified model is written as

\[
y = X_t w_s + u
\]

with polynomial restrictions being

\[
R_{w_s} = 0
\]

then generally the PDL estimator is biassed, and the general expression for this is

\[
E(\tilde{w}_s) = (I - M_s R_s)(X_t' X_t)^{-1} X_t' X_w.
\]

This result does not depend upon end-point restrictions.

III: \( n = n^*, p > p^* \)

**Proposition:** If the assumed polynomial has order higher than the true polynomial but the lag length is correct, the PDL estimator is unbiased but inefficient.

In this case the PDL estimator is\(^3\)

\[
\tilde{w}_0 = \tilde{w} - M_0 R_{p+1} \tilde{w}
\]

and its expectation is given by

\[
E(\tilde{w}_0) = E(\tilde{w}) - \sum_{k=p+2}^{r+1} A_k R_{p+1} E(\tilde{w}) = w
\]

where \( A_k \) is \((n - k) \times (n - k + 1)\) matrix such that \( R_{p+1} = A_{p+1} A_p \ldots A_1 \).

---

\(^3\) See footnote 2 for the definition of \( M_0 \).
There is an efficiency loss since it is clear that, although the two sets of restrictions are both correct, only \( n - p \) rather than \( n - p^* \) are actually imposed if \( p \) is different from \( p^* \), i.e., a smaller number of restrictions than those available is imposed whenever \( p > p^* \), thereby leading to an efficiency loss.

**IV: \( n = n^*, p < p^* \)**

**Proposition:** If the assumed polynomial has degree lower than the true polynomial and the lag length is correct, the PDL estimator is always biassed.

In this case the model \( y = Xw + u \) is estimated subject to incorrect restrictions \( R_u w = 0 \) where \( R_u \) is the appropriate sub-matrix of \( R \). The bias of the restricted least squares estimator, say \( \hat{w}_u \), is given by

\[
E(\hat{w}_u) = w - (X'X)^{-1}R_u'(R_u(X'X)^{-1}R_u')^{-1}w.
\]

3. **Choosing the Order of the Polynomial**

Because of the possibility of specification bias with an incorrect selection of \( n \) and \( p \), it is desirable to have some means for testing if the choice is “optimal” i.e., is the deterioration in the “goodness of fit” statistically significant if the order of the polynomial is changed from \( p \) to \( m \)?

To make any progress it is necessary to assume that \( n \) is known and that a choice of \( p \) needs to be made (although an identical procedure follows if one fixes \( p \) and varies \( n \), although an upper limit needs to be set to the range of \( n \)). Essentially the test statistics focus upon the \((n-p)\) linear restrictions imposed upon the coefficients, so that fixing either \( n \) or \( p \) allows the number of restrictions to change as the other is varied.

Fixing \( n \), the most general alternative hypothesis is that \( p = n \) i.e., no restrictions are applied, and the likelihood ratio test of the null hypothesis that \( p = m < n \) is based on the criterion

\[
Q_1 = \frac{(n - m)s^2}{\hat{\sigma}^2} - 1 \hat{\phi}'R'(R(X'X)^{-1}R')^{-1}R\hat{\phi}
\]

where \( s^2 \) is the OLS estimate of the residual variance. Under \( H_0, p = m, Q_1 \sim F_{n-m, T-n-1} \) enabling us to perform a test of the validity of the restrictions.

In the current situation though, there is a desire to test a sequence of hypotheses

\[
H_0: \quad p = n - 1
\]
\[
H_{n-1}: \quad p = n - 2
\]

etc.

i.e., beginning with no restrictions one is added at a time. Such a sequence is uniquely ordered and nested in that, if \( H_k \) is true, \( H_{k+j} (j > 0) \) must be true, and the process of testing \( H_n, H_{n-1} \) etc. should continue until some \( H_k \) is rejected. By appeal to the Independence Theorem due to Basu [7] and Hogg and Craig [15], it can be established that the significance level of the test of \( H_{n-q} \) against \( H_n \) is

\[
1 - \prod_{i=1}^{q} (1 - \gamma_i)
\]

where \( \gamma_i \) is the significance level chosen at the \( i \)th step in the sequence. If \( \gamma_1 = \gamma_2 = \cdots = \gamma_q \), the probability of rejecting the null hypothesis when it is true will rise as the order of the polynomial is reduced provided \( \gamma_i > 0 \), and hence also when \( \gamma_i = \gamma \) all \( i \). In his discussion of a similar problem Anderson [6, pp. 34-43] discusses the properties of such a procedure. He observes that “in general one has to balance the desirability of not over-estimating the degree (of the poly-
nomial) with the sensitivity of the procedure to non-zero coefficients”. In following this approach one varies the significance level with the value of $p$, making $\gamma_j$ very small for high values of $p$, so that if a high order polynomial is required there is a chance of learning that fact and, if not, the probability of deciding on a high order is small. An alternative along the lines suggested by Anderson would be to pick $p$ close to $n$ and let $\gamma_j$ increase with $j$ close to $n$ e.g. as in Anderson [6, p. 42], $\gamma_j = \gamma(n + 1 - j)/n - m, j = m + 1, \ldots, n$. See also Mizon [18].

4. Another Statistic for Determining Polynomial Order

Previous discussion has concentrated upon the determination of polynomial order by the compatibility of various differencing restrictions with the data, but there is nothing unique about this set of restrictions, as any non-singular transformation of them will yield an alternative set of restrictions with the same $Q_1$ statistic. In this section we investigate such an alternative—that proposed by Godfrey and Poskitt [13].

Combining (1) and (2) gives

$$y_t = \sum_{k=0}^{p} \alpha_k z_{k,t} + u_t$$

where $z_{k,t} = \sum_{j=0}^{n} j^k x_{t-j}$. From (17), if the degree of the polynomial were $p - 1$, the model would be

$$y_t = \sum_{k=0}^{p-1} \alpha_k z_{k,t} + u_t$$

making $z_{p,t}$ the extra regressor when the degree of polynomial is raised from $p - 1$ to $p$. In general, raising the degree of polynomial from $m$ to $n$ adds the regressors $z_{m+1,t}, \ldots, z_{n,t}$ so that a test of the hypothesis that $p = m$ is available by testing whether $\alpha_{m+1}, \ldots, \alpha_n$ are jointly zero.

In matrix terms we have

$$y = XJ\alpha + u$$

so that

$$\hat{\alpha} = (J'X'XJ)^{-1}J'X'y$$

and

$$\text{cov}(\hat{\alpha}) = \sigma^2(J'X'XJ)^{-1}.$$ 

Letting $H$ be the $(n - m) \times (n + 1)$ matrix $[0, I_{n-m}]$, the requirement that $\alpha_{m+1}, \ldots, \alpha_n$ be equal to zero leads to the linear restrictions $H\alpha = 0$, and these may be tested for validity with the test statistic

$$Q_2 = ((n - m)s^2)^{-1}\hat{\alpha}'H'(HJX'XJ)^{-1}H\hat{\alpha}$$

which is $F_{n-m, T-n-1}$ under the null hypothesis. But, as the alternative is that $p = n$, $J$ is square and hence $\hat{\alpha} = J^{-1}\hat{w}$, i.e.,

$$Q_2 = ((n - m)s^2)^{-1}\hat{w}'H'(HJ^{-1}X'XJ^{-1}J')^{-1}H\hat{w}$$

$$= ((n - m)s^2)^{-1}\hat{w}'(J^{-1})'H'HJ^{-1}X'X^{-1}(J')^{-1}H'J^{-1}H\hat{w}$$

$$= ((n - m)s^2)^{-1}\hat{w}'\tilde{B}(\tilde{B}X'X)^{-1}B'\tilde{B}\hat{w}$$

where $\tilde{B}$, the $(n-m) \times (n-m)$ sub-matrix formed from the inverse of $J$, is easily seen to be Godfrey and Poskitt's $\tilde{B}$ i.e. the $Q_2$ test statistic is that employed by those authors.

Is there any reason to prefer $Q_1$ to $Q_2$? Theoretically they must be identical as both $R$ and $\tilde{B}$
have equal rank and both are Wald test statistics, indicating that any choice must be made on numerical considerations. By presenting the derivation of $Q_2$ as above, one sees its affinity with the direct method of solving polynomial lag problems (to use Cooper's [8] terminology), and hence it is likely that the numerical difficulties of the direct method apply also to $Q_2$. It seems possible that $\hat{B}(X'X)^{-1}\hat{B}'$ will tend to be ill conditioned if $n$ is large and $m$ small since then it is necessary to reinvert “most” of $J^{-1}$ and the $|J^{-1}| = (n!(n-1)! \ldots 1)^{-1}$.

Later we examine the following Jorgenson-type investment equation—taken from Henderson and Norman [14].

$$I_t = \beta_1 + \beta_2 S_1 K_{E_{t-1}} + \beta_3 S_2 K_{E_{t-1}} + \beta_4 K_{E_{t-1}} + \sum_{j=0}^{n} \delta_j \Delta \left( \frac{\text{GDPS}}{p_e} \right)_{t-j}$$

where $I_t$ is gross private fixed capital expenditure on plant and equipment, $K_{E_{t-1}}$ is the lagged stock of plant and equipment, $S_i$ are seasonal dummies, GDPS is the value of gross domestic product and $p_e$ is the implicit rental rate of capital. In the paper by Henderson and Norman, $n = 17$, and our later analysis adopts this value, but for the purposes of the current experiment $n$ was varied from 8 to thirteen with the polynomial order ($p$) going from unity to the lag length in each case. Table 1 presents the values for $Q_1$ and $Q_2$, with $Q_1$ in brackets.4)

<table>
<thead>
<tr>
<th>Lag n/poly deg.</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.708</td>
<td>1.226</td>
<td>3.135</td>
<td>.468</td>
<td>9.704</td>
<td>-ve</td>
</tr>
<tr>
<td></td>
<td>(7.715)</td>
<td>(1.246)</td>
<td>(3.182)</td>
<td>(3.380)</td>
<td>(2.064)</td>
<td>(0.906)</td>
</tr>
<tr>
<td>2</td>
<td>1.701</td>
<td>3.307</td>
<td>2.870</td>
<td>12.123</td>
<td>.037</td>
<td>-ve</td>
</tr>
<tr>
<td></td>
<td>(1.700)</td>
<td>(3.299)</td>
<td>(2.817)</td>
<td>(12.118)</td>
<td>(3.105)</td>
<td>(3.688)</td>
</tr>
<tr>
<td>3</td>
<td>.094</td>
<td>.084</td>
<td>.567</td>
<td>.573</td>
<td>6.659</td>
<td>-ve</td>
</tr>
<tr>
<td></td>
<td>(.094)</td>
<td>(.083)</td>
<td>(.573)</td>
<td>(.604)</td>
<td>(.711)</td>
<td>(.047)</td>
</tr>
<tr>
<td>4</td>
<td>.026</td>
<td>.435</td>
<td>.009</td>
<td>.208</td>
<td>.107</td>
<td>5.817</td>
</tr>
<tr>
<td></td>
<td>(.026)</td>
<td>(.435)</td>
<td>(.009)</td>
<td>(.204)</td>
<td>(.757)</td>
<td>(.979)</td>
</tr>
<tr>
<td>5</td>
<td>.109</td>
<td>.018</td>
<td>.108</td>
<td>1.101</td>
<td>.984</td>
<td>.608</td>
</tr>
<tr>
<td></td>
<td>(.109)</td>
<td>(.018)</td>
<td>(.108)</td>
<td>(1.098)</td>
<td>(1.177)</td>
<td>(.022)</td>
</tr>
<tr>
<td>6</td>
<td>.239</td>
<td>.045</td>
<td>.095</td>
<td>.655</td>
<td>.318</td>
<td>.764</td>
</tr>
<tr>
<td></td>
<td>(.239)</td>
<td>(.045)</td>
<td>(.095)</td>
<td>(.655)</td>
<td>(.290)</td>
<td>(.546)</td>
</tr>
<tr>
<td>7</td>
<td>.039</td>
<td>.268</td>
<td>.579</td>
<td>.315</td>
<td>.109</td>
<td>1.746</td>
</tr>
<tr>
<td></td>
<td>(.039)</td>
<td>(.268)</td>
<td>(.579)</td>
<td>(.315)</td>
<td>(.109)</td>
<td>(1.691)</td>
</tr>
<tr>
<td>8</td>
<td>.124</td>
<td>.052</td>
<td>.333</td>
<td>.167</td>
<td>.029</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.124)</td>
<td>(.052)</td>
<td>(.333)</td>
<td>(.167)</td>
<td>(.029)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.138</td>
<td>.065</td>
<td>.935</td>
<td>.052</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.138)</td>
<td>(.065)</td>
<td>(.935)</td>
<td>(.052)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.251</td>
<td>.088</td>
<td>.199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.251)</td>
<td>(.088)</td>
<td>(.199)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>.129</td>
<td>.499</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.129)</td>
<td>(.499)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>.120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.120)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4) Data is available on request from the authors. All computations of $Q_2$ were performed with a program supplied by Leslie Godfrey, except that all variables in the program were double precisioned.
It is possible to verify that $Q_1$ is the correct value for the statistic by performing Almon regressions with all the $(n, p)$ combinations in the above table and constructing the Wald test statistics from the resulting sum of squares. By concentrating upon $n = 8, \ldots, 13$ and $p = 1$ it is possible to see how slight inaccuracies finally become major ($n = 12$) until there is a complete breakdown for $n \geq 13$ (the determinant of $\hat{B}(X'X)^{-1}\hat{B}'$ was negative for $n = 13$ and $p \leq 3$). The first serious breakdown in $Q_2$ therefore occurs at $n = 11$ but, if one reformulates the equations with levels of $\frac{\text{GNP}}{p_c}$ rather than differences, the breakdown occurs at $n = 9$ i.e., the accuracy depends on the amount of the collinearity in the data matrix. Therefore, although there are likely to be many cases when $Q_2$ will be accurate e.g. the short lags and non-collinear spirits data used by Godfrey and Poskitt, the possibility of failures leads one to select a more reliable method of testing for polynomial order i.e., $Q_1$.

5. Wider Considerations Relating to the Application of the Polynomial Model

The assumption that distributed lag weights lie on a polynomial is not intended to be taken as literally true; rather it is a simplifying assumption. Viewing polynomial selection in this fashion highlights the fact that concern over polynomial order may be misplaced (concern over lag length likely to be more meaningful as it is properly part of model specification). If one does decide that the polynomial approach is merely a convenient fiction, there are several avenues open to the empirical investigator:

1. Find a better story i.e., generalize the functional form. As Taylor shows [27], Shiller’s [25] smoothness prior estimator may be viewed as imposing stochastic differencing restrictions upon the parameters, while the method of spline lags advanced by Poirier [20] may be regarded as imposing non-stochastic linear restrictions, which are composed of differencing restrictions for the various spliced polynomials and others that represent continuity at the “knots”. One can also consider a more general form of the differencing restrictions, viz.,

$$(1 - L)^d w_j^\lambda \frac{-1}{\lambda} = 0,$$

where $\lambda$ is an unknown parameter. When $\lambda = 1$, imposing these restrictions would yield Almon’s estimator, while $\lambda = 0$ gives $(1 - L)^d \log w_j = 0$, i.e., Shiller’s recent suggestion [26]. Furthermore, because the above restriction implies that

$$w_j^\lambda = \alpha_0 + \alpha_j j + \cdots + \alpha_p j^p$$

one can even envisage the $\alpha_j$’s and $j$’s raised to powers of $\lambda$ as well. Clearly there are many possibilities of freeing the functional form, all sharing the dual disadvantage of arbitrariness and the necessity of adopting iterative techniques for their solution;

2. A Bayesian approach which uses prior information about lag coefficients and/or the lag length becomes an attractive alternative once the BLUE criterion is not regarded as compelling (see Leamer [16] and Mouchart and Orsi [19]). Here we briefly review the relationships between several estimators which are or can be thought of as based on Bayesian considerations, e.g. the James-Stein estimator, the ridge estimator and Shiller’s estimator. Conveniently assume that $y \sim N(Xw, \sigma^2 I)$. For given $n$ and $q$ let the prior distribution of $w$ be Normal, i.e.
\[ \pi(w | \sigma^2) \propto \sigma_w^{-(q+1)} \exp \left\{ -(1/2\sigma_w^2)(w' Dw) \right\} \]

where \( \sigma_w^{-2} \) is the prior precision parameter and \( D \) is a positive-definite matrix of order \((n + 1)\).

The mean of the posterior distribution of \( w \), for three alternative choices of \( D \), is as follows:

(a) \( D = R'R \) which leads to
\[ w_{Sh} = [X'X + (\sigma^2/\sigma_p^2)R'R]^{-1}X'y, \]

(b) \( D = I \) which leads to
\[ w_R = [X'X + (\sigma^2/\sigma_p^2)I]^{-1}X'y, \]

(c) \( D = XX' \) which leads to
\[ w_{JS} = (1 + k)^{-1}(X'X)^{-1}X'y \]

where \( k = b/w'w \) and \( b \) is a scalar. These can be readily identified as the Shiller, the ridge and the James-Stein estimators. Each may be thought of as a matrix—weighted average of the least-squares estimator \((X'X)^{-1}X'y\) and the mean of the assumed prior distribution of \( w \). The chief difference between the three lies in the assumption about the prior mean. Whereas for \( w_R \) and \( w_{JS} \) the prior mean is a vector of zeroes, in case of \( w_{Sh} \) only the mean of the \((q + 1)\)th order differences of \( w \) is zero so that the estimates are smoothed only and not pulled towards zero as in the case of the other two. Since sample information may quite unambiguously indicate that particular components of \( w \) are non-zero, uniform shrinkage towards zero seems not preferable except possibly in case of the weights in the tail of the distribution. The smoothness prior used by Shiller, therefore, appears to provide the most attractive of the Bayesian alternatives. All three estimators pose the problem of choosing the prior parameters, which have been assumed known in the above account. This problem however is outside the scope of this paper.

(3) Accept a more limited decision theoretic framework specifying a loss function and choosing a specification which is optimal relative to it. This last approach is examined in Section 6 and 7.

6. Estimation Under Quadratic Loss

The propositions discussed in section 2 would be of special interest only if there was a reasonable probability of including the true model in the set of specified models. Given such a possibility Theil’s \( \bar{R}^2 \) criterion is available for choosing a model and was the strategy selected by Schmidt and Waud [22]. Giles and Smith [12] have shown that, on average, this will select the correct \( n \) and \( p \) but, as the criterion is known to be defective in the case of the deletion of regressors where it yields the rule to delete if the corresponding F statistic is greater than unity, it would seem more appropriate to maximize \( R^2 = 1 - (1 - R^2)\left(\frac{T}{T-K}\right)^8 \) which Deaton [9] has shown to delete regressors whenever \( F > a \) and the ratio of the number of deleted regressors to degrees of freedom is small. If, however, it is accepted that the choice of integer \( p \) is only a matter of approximation, then the usual BLUE criterion loses its appeal, and an alternative, such as the quadratic loss criterion, which emphasises (for example) the predictive performance of the model appears more appropriate. Furthermore, it is the case that most of the final estimates of the Almon models used in applied work are obtained after extensive testing with values of \( n \) and \( p \), so that the estimators suffer from pre-test bias, see Frost [11]. This agains invalidates the use of the BLUE criterion. It is appropriate then to consider the consequences of mis-specification and the
properties of the estimator when the BLUE criterion is abandoned and the quadratic loss criterion substituted in its place.

Consider estimation relative to the loss function \((w - \hat{w})'X'X(w - \hat{w})\). Under quadratic loss a number of estimators of \(w\), e.g., the ridge estimator and the Stein-James estimator (as modified by Sclove [23]), are likely to be superior to the least squares estimator. It can also be shown that Shiller’s estimator [25] is related to the ridge-type estimators and hence will share some of their desirable properties, though it is of course motivated by considerations more general than minimizing mean square error. Such estimators have not been widely used in empirical work—an exception is Aigner and Judge [1]—but there seems to be a good case for so doing in the present context.\(^5\)

An alternative approach is to consider what losses may arise from the mis-specification discussed in previous sections. Amemiya and Morimune (A—M) [4], by assuming that the independent variable is generated by a first-order stationary autoregressive process with \(E_{x_t} = 0\), \(E_{x_t x_{t+n}} = \sigma^2 x^2 (p^2 / 1 - \rho^2)\), have considered the following loss function

\[
l^*(\hat{w}) = (\sigma^2 x^2 w' \Lambda w)^{-1} \left[ \text{tr MSE}(\hat{w}) X'X / T \right]
\]

where \(\sigma^2 x^2 w' \Lambda w\) is a normalizing constant and

\[
\Lambda = (1 - \rho^2)^{-1} \begin{pmatrix}
1 & \rho & \cdots & \rho^n \\
\rho & & & \\
\vdots & & & \\
\rho^n & & & 1
\end{pmatrix}
\]

The rationale for this loss function has been given in A—M and will not be repeated here. For the Almon model A—M show that the loss can be expressed as

\[
l^*(w) = T^{-1}(\sigma^2 x^2 w' \Lambda w)^{-1} \sigma^2 (p + 1) + (w' \Lambda w)^{-1} (w' R'(R \Lambda^{-1} R')^w w)
\]

and this enables the tabulation of the losses due to all the mis-specifications previously mentioned (see Table 2). To interpret Table 2 the following additional notation is required

\[
R_A = \begin{pmatrix}
R_S & 0 \\
0 & I
\end{pmatrix}, \quad \Lambda_A = \begin{pmatrix}
\Lambda & \Lambda_{12} \\
\Lambda_{21} & I
\end{pmatrix}
\]

\(I\) is an identity matrix of rank \(n - n^*\).

From these expressions it can be seen that

\[
l^*(\hat{w}) \leq l^*(\hat{w}) \quad \text{as } p^* \leq n^*,
\]

\[
l^*(\hat{w}) \leq l^*(\hat{w}_L),
\]

\[
l^*(\hat{w}) \leq l^*(\hat{w}_L),
\]

\[
l^*(\hat{w}) \leq l^*(\hat{w}_0).
\]

Only in the case of \(\hat{w}_u\) is it not possible to make an unambiguous comparison between \(l^*(\hat{w})\) and \(l^*(\hat{w}_u)\). This is rather unusual in that no such unambiguous comparison could be made between

---

5) Aigner and Judge [1] have pointed out that the necessary condition for certain Stein-James estimators to dominate the OLS estimator (given a quadratic unweighted loss function \((\hat{w} - w)'(\hat{w} - w)\)) is that \(\text{tr} \frac{(X'X)^{-1}}{\lambda} > 2\) where \(\lambda\) is the largest latent root of \((X'X)^{-1}\). With multicollinear data sets this inequality may not be satisfied.
Table 2 Losses Under Different Mis-specifications

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Normalised Loss Function: ( l^*(\cdot) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\varphi} )</td>
<td>( T^{-1} \sigma^2(n^* + 1)/[\sigma_x^2(w'Aw)] )</td>
</tr>
<tr>
<td>( \hat{\varphi}_n )</td>
<td>( T^{-1} \sigma^2(p^* + 1)/[\sigma_x^2(w'Aw)] )</td>
</tr>
<tr>
<td>( \hat{\varphi}_L )</td>
<td>( T^{-1} \sigma^2(p^* + 1)/[\sigma_x^2(w'Aw)] )</td>
</tr>
<tr>
<td></td>
<td>+ ( w'R_{1-1}R_{1-1}^{-1}w/\sigma^2 ) ( \quad (n &gt; n^*) )</td>
</tr>
<tr>
<td>( \hat{\varphi}_0 )</td>
<td>( T^{-1} \sigma^2([\sigma_x^2(w'Aw)][(p^* + 1)][(p^* - p) + 1]) )</td>
</tr>
<tr>
<td>( \hat{\varphi}_w )</td>
<td>( T^{-1} \sigma^2([\sigma_x^2(w'Aw)][(p^* - p) - (p^* + 1)] )</td>
</tr>
<tr>
<td></td>
<td>+ ( w'R_{2-1}R_{2-1}^{-1}w/\sigma^2 ) ( \quad (p &lt; p^*) )</td>
</tr>
</tbody>
</table>

\( \hat{\varphi}_n \) and \( \hat{\varphi}_0 \) if our loss function was based strictly on MSE; whether this tells us something about the impossibility of gains by understating lag length or the inadequacies of the A—M type of loss function depends upon the user’s preferences. If one were happy with A—M’s formulation, it appears that mis-specification can only lead to reduced loss if it involves an underatement of polynomial order—selecting short lags will not be sufficient.

Another approach which does not seem to have been tried out systematically in the estimation of polynomial distributed lags, but which has potential applicability is Akaike’s information criterion, proposed by Akaike [2] and discussed by Amemiya [5]. This criterion or the prediction error criterion of Amemiya [5] can be used in selecting the number of regressors as well as in choosing among various linear constraints.

7. Restricted Estimation

In most applied situations characterized by long lags it is usual to reduce considerably the dimensionality of the problem by imposing a number of linear restrictions. For example, \((n - p) \leq 12 \) and \( 15 \) are not uncommon. This could be justified by appeal to either an unweighted quadratic loss criterion \( \sum (w_i - \hat{\varphi}_i)^2 \), or a weighted quadratic loss criterion \( (\hat{\varphi} - w)'X'X(\hat{\varphi} - w) \), for any estimator \( \hat{\varphi} \), or the minimum average relative risk criterion\(^6\) put forward by Toyoda and Wallace [29] in their extension of Sawa and Hiromatsu [21]. The precise choice depends upon the objective of the investigator, and in particular, the second criterion seems appropriate in a prediction context where consideration of overall gain is more important.

---

6) Let \( \rho(\beta^*, \beta) \) denote the risk of the pre-test estimator \( \beta^* \) such that

\[
\beta^* = \begin{cases} 
\hat{\beta} & \text{if } u \geq \lambda \\
\hat{\beta} & \text{if } u < \lambda
\end{cases}
\]

where \( \hat{\beta} \) is the unrestricted least squares estimator, \( \hat{\beta} \) is the restricted least squares estimator, \( u \) is the test statistic distributed as a noncentral \( F \) variate with noncentrality parameter \( \theta \) and \( \lambda \) the critical value for accepting or rejecting some hypothesis about \( \beta \). Relative risk of \( \beta^* \) relative to \( \beta \) and \( \hat{\beta} \) is defined as

\[
\rho(\beta^*, \beta) = \min_{\beta} \left[ \rho(\beta^*, \beta), \rho(\hat{\beta}, \beta) \right] / \sigma^2
\]

and the minimum average relative risk criterion purports to minimise this criterion over the whole range of \( \theta \).
than the cost of worsening an individual estimate. Of course, the three criteria need not necessarily lead to the same choice of restrictions so it is interesting to consider their implications for the choice of \( p \).

Using a minimum average relative risk criterion they show that this risk function for the pre-test estimator "leads to opting for ordinary, unrestricted least squares unless the number of restrictions is five or greater, no matter the number of regressors and observations, i.e. our optimal critical value of pre-test is zero unless the restrictions imply a reduction of five or more in the parameter space". When the number of restrictions increases to (say) eight, the optimal critical value of the pre-test estimation increases to 1.059, 1.170 and 1.215 as the degrees of freedom increases to 24, 60, 120 respectively. In the case of polynomial distributed lag models, the implication is that \( p \leq n - 5 \) and that the reduction in \( p \) should take place only if the test statistic is less than the critical value. Thus the minimum average relative risk criterion gives some positive chance to the polynomial formulation being accepted.

However, it is known (see Sclove [24]) that variants of the Stein-James estimators which combine the restricted and unrestricted estimates dominate the pre-test estimators for the minimum average relative risk criterion and this suggests that one might do better adopting such an estimator.

To some extent, however, the above discussion is a little artificial in that we consider the problem as one of optimally combining restricted and unrestricted estimators when in practice investigators often consider a sequence of restrictions before adopting any one set in such a sequence. What we need therefore is an estimator which combines the information contained in a sequence of pre-test estimators. A remarkable recent result due to Learner and Chamberlain [17, pp. 86–88] establishes a connection between \( 2^k \) restricted estimates of a \( K \)-component regression coefficient vector and the Bayes posterior mean. They show that "the posterior mean is a convex combination of the estimates resulting from all combinations of dropping and not dropping variables". This result suggests that reporting the Bayes posterior mean may not be a great deal less informative and may be more informative than reporting a sub-set of results thought by an investigator to be interesting or relevant.

**An Illustration**

The investment equation previously discussed is now reconsidered with a view to illustrating a method of choice between the restricted, the unrestricted and the modified positive part Stein-James rule due to Sclove et al. [24]. The restrictions are those that arise from the choice of a second-degree polynomial to generate the lag-weights; the reason for such a choice being that this is what one would select on the basis of a sequential test using \( Q_1 \) if \( \gamma = 0.05 \) for all \( i \) and it

---

7) The modified positive part role in the present context is

\[
w^{**} = I_{(c, \infty)}(u)(1 - c^*/u)^* \hat{\beta}
\]

where \( I_{(c, \infty)} = 1 \) if the test-statistic \( u \) is in the interval \([c, \infty)\), \( 0 < c^* \leq 2(K - 2)(T - K)/(T - K + 2)K \), \( c \) being the critical value of the \( F \)-statistic at a prechosen significance level, and \( K \) the total number of regressors in the equation.
is also a polynomial order frequently selected by investigators.\(^8\)

The first question is whether one would accept a second degree polynomial on the relative average risk criterion—this being answered in the affirmative by comparing \(Q_1 = .2732\) to the optimal critical value of 1.381 provided by Wallace and Toyoda for the number of restrictions and degrees of freedom being 16 and 16 respectively (no values are available for our values of 15 and 19 but the test would be unlikely to yield a different result).

If it is now assumed that the investigator's main interest is in prediction then the modified positive part rule is applied as follows:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c \leq \frac{2(K - 2)(T - K)}{(T - K + 2)K})</td>
<td>Set (c^* = c; \bar{w}^* = \left(1 - \frac{c^*}{u}\right)\hat{\psi})</td>
</tr>
<tr>
<td>(c &gt; \frac{2(K - 2)(T - K)}{(T - K + 2)K})</td>
<td>(w^* = \begin{cases} \bar{\psi} &amp; \text{if } u \leq c \ \left(1 - \frac{c^*}{u}\right)\hat{\psi} &amp; \text{if } u &gt; c \end{cases})</td>
</tr>
</tbody>
</table>

Using the investment data it was found that with \(n = 17, q = 2, K = 22, T = 41, u\) took the value 0.2732. The five and one per cent critical values (\(c\)) are 2.23 and 1.86 respectively which exceed \(2(K - 2)(T - K)/(T - K + 2)K = 1.645\). Since \(u\) is less than \(c\), the application of the rule leads to the choice of second-degree polynomial estimates, i.e., the restricted least squares estimates.

Too much need not be read into the above illustration. Note that it does not use the variant of the James-Stein estimator which would uniformly (i.e. by the same factor) shrink the least squares estimates towards zero. Clearly this will never give exactly the same results as a restricted least squares estimator. The example simply elucidates why, in many applied situations, a restricted least squares estimator may be preferred. Furthermore, note that in many applied situations the predictive performance of the restricted least squares estimator may be quite hard to improve upon using variants of the James-Stein estimator (see Trivedi [31] esp. section 6), so there is some justification for choosing the restricted least squares estimator.

\((The\ Australian\ National\ University)\)

**REFERENCES**


\(^8\) It is worth noting that in their work on the optimal order of polynomial Amemiya and Morimune give several numerical examples where \(p\) increases discontinuously with the other parameters in the model (see A—M [4, pp. 381–382]).
April 1979  P.K. Trivedi and A.R. Pagan: Polynomial Distributed Lags


