TREATMENT OF LAGGED DEPENDENT REGRESSORS AND AUTOCORRELATED ERRORS IN SAS/ETS® SOFTWARE

Thomas B. Fomby and Joseph G. Hirschberg
Dept. of Economics, Southern Methodist University, Dallas, 75275

1. Purpose

The purpose of this paper is two-fold. First, we wish to critique the current documentation of the PROC AUTOREG procedure as described in the SAS/ETS User's Guide (Version 5), especially as it relates to the discussion of Lagged Dependent Regressors (pp. 192-3). In particular, the problem we are interested in is characterized by the regression model

\[ y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \nu_t \]  
\[ \nu_t = \epsilon_t - a_1 \nu_{t-1} - \cdots - a_p \nu_{t-p}, \quad t=2,...,T, \]  

where \( \epsilon_t \) is a sequence of independent normal error terms with mean zero and variance \( \sigma^2 \), \( x_t \) denotes a \((K-2) \times 1\) vector of observations on nonstochastic explanatory variables at time \( t \), \( |\beta_1| < 1 \) and the values of the \( a_i \)'s are such as to guarantee the covariance stationarity of the error process \( \nu_t \). The lagged dependent regressor - autocorrelated error model of (1) and (2) will often be referred to in compressed form as

\[ y = X\beta + \nu, \]  

where \( y = (y_2,...,y_T)' \), \( \beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_K)' \), \( \nu = (\nu_2,...,\nu_T) \) and

\[ X = \begin{bmatrix} 1 & y_1 & X_{1,1} & \cdots & X_{1,K-2} \\ 1 & y_2 & X_{2,1} & \cdots & X_{2,K-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & y_{T-1} & X_{T-1,1} & \cdots & X_{T-1,K-2} \\ 1 & y_T & X_{T,1} & \cdots & X_{T,K-2} \end{bmatrix} \]  

If the vector of autocorrelation parameters \( \gamma = (a_1, a_2, \ldots, a_p)' \)

were known, the variance-covariance matrix of \( \nu \), \( \Sigma = \sigma^2 \nu \), could be computed up to the scale factor \( \sigma^2 \). Efficient estimation of \( \beta \) and \( \sigma^2 \) then could proceed by means of generalized least squares (GLS). Unfortunately, \( \Sigma \) is usually unknown and, thus, alternative methods of estimation must be pursued. Atypically, estimated generalized least squares (EGLS) in this context is inefficient, see Maddala (1971) and Wallis (1972) on this matter.

The included sample program in the User's Guide (pp. 192-3) is intended to provide an estimation procedure for the model of equations (1) and (2). Unfortunately, the illustrated method appears to be inconsistent and thus asymptotically inefficient.

Our second purpose is to review some recent contributions to the literature in this area which go undocumented in the User's Guide. Emphasis is placed on the Hatanaka (1974) method and a first observation correction suggested by Harvey (1981). This method is asymptotically efficient and, because of the first observation correction, appears to exhibit good small sample performance as well.

The outline of the rest of the paper is as follows. In section 2 the SAS/ETS PROC AUTOREG procedure is examined for its applicability in the context of the model of equations (1) and (2). Also the User's Guide sample program is examined and the statistical properties of the recommended estimation procedure are discussed. In section 3 a review of the recent literature is presented with special emphasis on the Hatanaka procedure with first observation correction. A sample program is provided in the appendix (Figure 1) which executes this procedure.

2. Applicability of PROC AUTOREG

2.1 User's Guide Warning

The User's Guide does warn the SAS user that the PROC AUTOREG estimation procedures are not strictly applicable when lagged dependent regressors and autocorrelated errors coexist. The Guide states:

The Yule-Walker estimation methods are not directly appropriate for estimating models that include lagged dependent variables among the regressors. Also, in this case the maximum likelihood estimators will not be exact maximum likelihood estimators but will be maximum likelihood conditional on the first few values of the dependent variable (pp. 192-3).

The Guide does not, however, state why the Yule-Walker method is no longer applicable nor how appropriate maximum likelihood estimation might be in this case given the default starting values of the AUTOREG procedure.

2.2 Shortcomings of Yule-Walker Default

The reason that the Yule-Walker estimation method is not applicable (even indirectly) in
this context is because the structural (least squares) residuals used in the first-stage Yule-Walker equations are inconsistent. This inconsistency occurs because the least-squares estimates of the regression coefficient vector $\beta$ are themselves inconsistent due to the contemporaneous correlation that exists between the lagged dependent regressor and the autocorrelated errors of the regression model. Given the inconsistency of the least squares residuals, the initial estimates of the autoregressive error parameters $\alpha$ obtained from the Yule-Walker equations are likewise inconsistent. Since the Yule-Walker estimates of $\alpha$ are used in the EGLS estimation of $\beta$, the EGLS estimate of $\beta$ is likewise inconsistent.

Statistical inference on the regression coefficients based on the Yule-Walker method is likewise inconsistent. In the Yule-Walker method the estimate of the variance-covariance matrix of the coefficient estimates is taken to be $s^2 (X'V^{-1}X)^{-1}$ which is inappropriate. This estimate assumes that the corresponding population variance-covariance matrix is of the form $\sigma^2[I - (T-1)\alpha a']^{-1}$. The Aitken (GLS) variance-covariance matrix is not block diagonal according to partitions with respect to $\beta$ and $\alpha$. The Aitken variance-covariance matrix is not achievable even asymptotically.

To put this into concrete terms, consider the model of equations (1) and (2) again with $p = 1$. Dhrymes (1971, p. 201) has shown that the appropriate consistent estimate of the variance-covariance matrix of efficient estimates of $\beta$ is $\frac{s^2}{\sigma^2} (X'V^{-1}X - (T-1)\alpha a'a)^{-1}$, where $s^2$ and $\sigma^2$ are consistent estimates of $s^2$ and $\sigma^2$ respectively, $T$ is the sample size, $V$ is a consistent estimate of $V$ and $\alpha$ is a $K \times 1$ vector (K being the dimension of $\beta$) all of whose elements are zero except the second which is $\frac{s^2}{(1+\alpha a')}$, $\alpha$ being a consistent estimate of the regression coefficient associated with the lagged dependent variable regressor. Even if the appropriate formula for estimating the variance-covariance matrix of the efficient estimates of $\beta$ was recognized to be (6), substituting in the inconsistent estimates of $s^2$, $\sigma^2$, and $\beta$ provided by the default YW method would nonetheless lead to an inconsistent estimate.

Regardless, the statistical inferences provided by the t-statistics and standard errors reported by the YW method are inconsistent.

Finally, further iteration of the YW method is not necessarily going to improve matters as far as the efficiency of estimation is concerned. Additional iteration on two-step procedures are, in general, likely to lead to a globally false minimum.

2.3 Shortcomings of ML Estimation

A readily available alternative is the maximum likelihood (ML) method. The User's Guide warning quoted above does not portend any difficulties when using this method. Conditional on the initial values of the lagged dependent regressors, the ML estimates are computed by minimizing $\frac{1}{T} (y' - X\hat{\beta})^2$ where $y = (y - X\beta)$ denotes the errors of the regression model and $L$ is the Cholesky root of $V$ such that $V = LL'$. But one suspects that this objective function is multimodal (see Harvey (1981, p. 125)) and, as a result, inconsistent starting values for the parameters $\beta$ and $\alpha$ (as would be generated by the least squares and Yule-Walker methods) might lead to a local minimum being chosen by iteration methods rather than to a globally true minimum. Given the possibility of a local minimum solution persisting even in infinite samples because of inconsistent starting values, the ML algorithm as implemented by PROC AUTOREG may yield inconsistent and thus inefficient estimates of the parameters just as the default method (YW) will.

Let $J$ denote the Jacobian matrix of derivatives of $v$ with respect to the parameters evaluated at the ML estimates. Though the matrix $\frac{\partial^2 v}{\partial (\beta, \alpha, \sigma^2)}^{-1}$ is, in principle, an appropriate consistent estimate of the variance-covariance matrix of the parameter vector $(\beta, \alpha, \sigma^2)$, the fact still remains that the expression $\frac{\partial^2 v}{\partial (\beta, \alpha, \sigma^2)}^{-1}$ may be evaluated at inconsistent estimates obtained from a ML iterative procedure using inconsistent starting values and applied to a multimodal objective function.

2.4 User's Guide Literature Review

The literature references supplementing the discussion of the lagged dependent regressor case are at best incomplete. The first reference Johnston (1972, p. 319) contains a description of the Wallis (1972) two-step EGLS procedure. Though this method is consistent it
is not, however, asymptotically efficient. Moreover, naive use of the EGLS variance-covariance expression (rather than (6) above) will lead to inconsistent inference.

The Fuller (1976, pp. 429-447) reference recommends the use of the Gauss-Newton method. The initial estimate used in this two-step scheme is obtained by the method of instrumental variables. As proposed, the Fuller procedure is consistent as well as asymptotically efficient though, as will become apparent in section 3, the treatment of the first observation appears to be done in a way incongruent with the spirit of the Gauss-Newton method. Regardless, the treatment of the first observation is not crucial to the asymptotic efficiency of the proposed estimator but may affect its small sample efficiency.

2.5 User's Guide Sample Program

There is some question as to the appropriateness of the SAS program included in the User's Guide (p. 193) as a "solution" to the present estimation problem. The steps implemented by the program are: (1) Use least squares to obtain an instrument for the "troubling" variable $Y_{t-1}$. That is, regress $Y_t$ on $X_t$ and $X_{t-1}$ by means of least squares and retain $\hat{Y}_t$, the predicted values of $Y_t$. (2) Use PROC AUTOREG to regress $Y_t$ on the nonstochastic explanatory variables and $Y_{t-1}$ and at the same time correct for the first order autocorrelation in the errors by implementing the NLAG = 1 option.

A problem arises, however, in the simultaneous application of two-stage least squares (2SLS) and a first order autocorrelation correction via the PROC AUTOREG procedure. The default procedure here is YW. This means that the residuals of the second stage are those obtained from least squares applied to a "purged" equation, i.e. 2SLS residuals. Then these residuals are used to compute an estimate of $\rho$, the autocorrelation coefficient. But as Schmidt (1976, Proposition 12, p. 160) proves, the least squares residuals obtained from the "purged" equation are inconsistent. Instead consistent residuals should be obtained by computing the residuals from the original equation with $Y_{t-1}$, not $Y_{t-1}'$, as a regressor. Using the notation of (1) and (2), the consistent residuals are calculated as

$$ \hat{\epsilon}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t - \hat{\beta}_2 \hat{Y}_{t-1}, \quad (7) $$

where $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ represent the 2SLS estimates obtained by running least squares on the purged equation. We suspect that the residuals in the Sample Program are calculated as

$$ \hat{\epsilon}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t - \hat{\beta}_2 \hat{Y}_{t-1} - \hat{\epsilon}_t, \quad (8) $$

which are inconsistent. Use of the $\hat{\epsilon}_t$ residuals to calculate an estimate of $\rho$, to implement in a Prais-Winsten (Cholesky root) transformation will then necessarily lead to inconsistent estimation of $\hat{\beta}_0$, $\hat{s}_t$, and $\hat{\epsilon}_t$. Moreover, the YW default method does not consider the modification needed for the Atikven variance-covariance formula as described in equation (6) above.

3. Hatanaka's Procedure with First Observation Correction

Before moving to an estimation method we would recommend, a brief review of some of the literature on the present problem will be useful. Maddala (1971) and Wallis (1972) were first to note that two-step EGLS methods applied to this estimation problem are generally inefficient. Hatanaka (1974) was the first to propose a two-step procedure (not of the EGLS class) which is both consistent and asymptotically efficient. As proposed, Hatanaka did not specify a first observation correction for his procedure.

Harvey (1981, pp. 269-70) clearly identified Hatanaka's procedure as being algebraically equivalent to the Gauss-Newton method. This identification allowed him to propose a suitable first observation correction for Hatanaka's procedure. Monte Carlo investigations by Fomby and Guilkey (1983) and Fomby (1986) indicate that substantial gains in small sample efficiency can be obtained by correcting the first observation in Hatanaka's procedure.

To formalize our discussion and for notational convenience, the model to be considered is

$$ Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \quad (9) $$

$$ \epsilon_t = \rho \epsilon_{t-1} + \epsilon_t \quad (10) $$

where $X_t$ is nonstochastic, $|\beta_1| < 1$, $|\rho| < 1$ and the $\epsilon_t$ are independent and identically distributed normal random errors having zero mean and finite variance $\sigma^2$, the initial value of $\epsilon_0$ being realized in the infinite past.

The Gauss-Newton method is especially useful when ML estimation (or its asymptotic equivalent) proceeds by minimizing a sum of squares function, say $S(\psi) = \sum \epsilon_t^2$, where $\epsilon_t = \epsilon_t(\psi)$ is an "error" which depends on the value taken by $\psi$. The two-step Gauss-Newton estimator of $\psi$, say $\hat{\psi}$, is obtained by computing

$$ \hat{\psi} = \arg \min S(\psi) = \arg \min \sum \epsilon_t^2, \quad (11) $$

where $\epsilon_t = \epsilon_t(\psi)$.
where \( \hat{\psi} \) is an initial consistent estimate of \( \psi \) and the gradient vector \( (\partial \psi_t / \partial \psi) \) and \( \epsilon_t \) are evaluated at \( \hat{\psi} = \hat{\psi} \). Should the minimization of \( S(\psi) \) with respect to \( \psi \) provide asymptotically efficient estimates of \( \psi \), the Gauss-Newton estimator, \( \hat{\psi} \), will likewise be asymptotically efficient.

With respect to the lagged dependent regressor-autocorrelated errors model of (9) and (10), let

\[
\epsilon_t = (1-p^2)^{1/2}y_{t-1}/(1-p^2)^{1/2}y_{t-1}/(1-p^2)^{1/2}x_{t-1} \quad (12)
\]

and

\[
\epsilon_t = (y_t - \hat{\beta}_1 x_{t-1}) - \hat{\beta}_1 (y_{t-1} - \hat{\beta}_1 x_{t-1}) - \hat{\beta}_2 (x_{t-1} - \hat{\beta}_2) \quad (13)
\]

\( t = 3, 4, \ldots, T \).

Harvey (1981, pp. 259-70) has shown that the application of the two-step Gauss Newton method to the sum of squares function

\[
S_0(\psi) = \sum_{t=3}^{T} \epsilon_t^2 \quad (14)
\]

where \(\bar{\psi} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\rho})\), is algebraically equivalent to Hatanaka's estimator. Note the first observation is not treated here. In light of this correspondence, Harvey (1981, pp. 270-1) suggested a treatment of the first observation by applying the two-step Gauss-Newton method to the sum-of-squares function

\[
S(\psi) = \sum_{t=2}^{T} \epsilon_t^2 \quad (15)
\]

The following process computes the Hatanaka estimates of \( (\hat{\beta}_1, \hat{\beta}_2, \hat{\rho}) \) using a first observation correction in the model (9) and (10). We will present the steps in algebraic form and refer to the SAS language statements used to perform the estimation as given in figure 1 in the appendix.

1) Compute a vector of predicted values of \( Y \) using the lagged values of the nonstochastic explanatory variables and a parameter vector estimated in the regression:

\[
Y_t = \hat{\gamma}X_{t-1} + \epsilon_t \quad (16)
\]

thus the predicted values are:

\[
Y_t = Y_{t-1} + \epsilon_t \quad (17)
\]

where \( \hat{\gamma} \) is the LS estimate of \( \gamma \).

In SAS we use PROC REG and save the predicted values.

2) Using the purged \( Y_{t-1} \) in place of \( Y_{t-1} \) in the estimation of the model (9), we have the "purged" equation

\[
y_t = \hat{\beta}_1 y_{t-1} + \hat{\beta}_2 x_{t-1} + \epsilon_t \quad (18)
\]

Figure 1 shows the use of the predicted value of \( Y_{t-1} \) in a regression using PROC REG and saving the coefficient estimates.

3) Now the consistent estimates of \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) obtained from equation (17) are used to compute the corresponding consistent estimates of the residuals \( \hat{\epsilon}_t \).

Note this is the step that cannot be performed by simply applying the Yule-Walker estimation method to equation (17) via SAS PROC AUTOREG because the residuals will be computed using \( Y_{t-1} \) instead of the actual \( Y_{t-1} \). Thus we compute:

\[
\hat{\epsilon}_t = y_t - \hat{\beta}_1 y_{t-1} - \hat{\beta}_2 x_{t-1} \quad (19)
\]

where \( \hat{\beta}_1, \hat{\beta}_2 \) are the estimates from equation (18).

In SAS we use PROC SCORE with a name change for the parameter on \( Y_{t-1} \) to match it with the data for the actual value of \( Y_{t-1} \).

4) Compute the autoregressive parameter \( \hat{\rho} \) in equation (10) with the following regression, the estimated residuals taken from step 3.

\[
\hat{\epsilon}_t = \hat{\rho} \hat{\epsilon}_{t-1} + \eta_t \quad (20)
\]

In SAS it is necessary to create a lagged predicted residual from the output from PROC SCORE in a DATA statement. Then a regression is run on the residuals using PROC REG with the NOINT option on the MODEL statement and returning the value of \( \hat{\rho} \) in a data set being sure to rename the parameter to RHO because we need to refer to the values of \( \hat{\epsilon}_t \) as well.

5) Now transform the regressors in model (9) using the Prais-Winsten transformation for the first observation, but also include the
estimated residuals \( \hat{v} \) to create the following set of regressors.

\[
X^* = \begin{bmatrix}
(1-p^2)^{1/2}Y_1 & \cdots & Y_{T-1}(1-p^2)^{1/2}T-2 \\
(1-p^2)^{1/2}X_2 & \cdots & X_{T-1}(1-p^2)^{1/2}T-2 \\
\vdots & \ddots & \vdots \\
(1-p^2)^{1/2}v_2 & \cdots & v_{T-1}(1-p^2)^{1/2}T-2
\end{bmatrix}
\]

\( y^* = \begin{bmatrix}
(1-p^2)^{1/2}Y_1 \\
(1-p^2)^{1/2}X_2 \\
\vdots \\
(1-p^2)^{1/2}v_2 \\
v_2 \\
\vdots \\
v_{T-1}
\end{bmatrix}
\]

(21)

In SAS we perform this transformation in two separate DATA creation subprograms because of the need for the LAG of a value for a subset of the observations.

6) Now the Hatanaka estimates of \( \hat{\beta_1} \) and \( \hat{\beta_2} \), say \( \tilde{\beta}_1 \) and \( \tilde{\beta}_2 \), using a first observation correction, are obtained by computing a least squares regression of \( y^* \) on \( X^* \):

\[
\begin{bmatrix}
\tilde{\beta}_1 \\
\tilde{\beta}_2 \\
\rho
\end{bmatrix} = (X^* X^*)^{-1} X^* y^*
\]

(23)

The Hatanaka estimate of \( \rho \), say \( \hat{\rho} \), using a first observation correction is \( \rho = \tilde{\rho} \times \tilde{\rho} \).

In SAS the procedure we need to include a transformed intercept as well in a PROC REG statement for these final estimates.

This concludes the procedure.

The estimator originally proposed by Hatanaka (1974) (which did not implement a first observation correction) is obtained by repeating the steps above except the first observation in the matrices \( y \) and \( X \) is omitted. In passing, note that the first observation correction for the "residual adjustment" regressor consists of the term \( \bar{p} v_2/(1-p^2)^{1/2} \). This adjustment is derived by exploiting the Gauss-Newton method correspondence pointed out by Harvey (1981). In contrast, following Fuller (1976, p. 340), the correction he would propose for the present model is \( (1-p^2)^{1/2}v_2 \). Fuller’s procedure is likewise developed via the Gauss-Newton method. This is the anomaly that we earlier mentioned about Fuller’s first observation correction.

It should also be noted that the conventional ordinary least squares formula for the variance-covariance matrix of the coefficient estimates is correct when applied to the "residual adjusted" transformed data of (21) and (22). That is, if one wanted the consistent estimates of the variances and covariances of \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \), then the upper submatrix of \( \hat{\sigma}^2(X'X)^{-1} \), where \( \hat{\sigma}^2 = (y^* - X^* \hat{\beta}_1)'(y^* - X^* \hat{\beta}_1)/T \) and \( \hat{\beta}_1 = (\hat{\beta}_1, \hat{\beta}_2)' \), would be appropriate.

Simulations by Fomby (1986) indicate that substantial efficiency gains are obtainable from the first observation correction described above. Moreover, the Hatanaka procedure with first observation correction performed equally as well as a full ML grid search in the simulations conducted therein.

Given the above considerations, our recommendation is that the Hatanaka method with first observation be used when estimating models with lagged dependent regressors and first order autocorrelated errors. The Gauss-Newton is likewise applicable when the order of the autocorrelation in the errors is greater than one.

References


Figure 1. The Use of Standard SAS PROCs to Compute the Hatanaka Estimator Employing the First Observation Correction.

Routine to compute the Hatanaka estimator using SAS PROCs

* The non-stochastic regressors (X)
* The transformed X
* The lagged X
* The dependent variable (Y)
* The transformed Y
* The lagged Y
* The transformed lagged Y

Run OLS on the lagged dependent variable to create a predicted value for the lagged dependent variable.

PROC REG; MODEL YL_i .. XL_i ;

Estimate the model using the predicted lagged dependent variable as the regressor.

PROC REG OUTEST=CONEST (RENAME=(PYL-YL_M)) ;
RES:MODEL Y .. X .. PYL ; OUTPUT OUT = NEW_R_RES;

Use PROC SCORE to compute the consistent residuals for use in the estimation of RHO.

PROC SCORE DATA = NEW SCORE = CONEST OUT = NEWR TYPE = OLS ;
VAR Y .. X .. YL_M ;

Set the lag of the residuals to be used as the dependent term. Recall the residual is the name of the model (RES) and it is the negative residual.

DATA NEWR ; SET NEWR ; LR .. LAO(-RES) ; RES = -RES ;

Run a regression through the origin of the residuals on the lagged value.

PROC REG OUTEST=NEST(RENAME = RES-RHO) ;
MODEL LR .. RES / NOINT ;

Change the name of the coefficient to RHO then add it to the data. Setup new values of the dependent and independent variables that are transformed with these values of RHO, (all but the first value). INT is the intercept term that is now transformed.

DATA NEW1 ; MERGE NEW NEWR(KEEP .. RES) ; INT .. 1 ; RETAIN RHO ;
IF _N_ = 1 THEN SET NEST(KEEP .. RHO) ;
ARRAY XX Y .. INT X .. YL_M ; ARRAY MX RTX_M RTY_M KYL_M ;
DO OVER XX ; MX = XX - RHO * LAG1(XX) ; END;
PROC PRINT ;
DATA NEW1 ; SET NEW1 ;

Transform the first observations.

ARRAY XX Y .. INT X .. YL_M ; ARRAY MX RTX_M RTY_M KYL_M ;
IF _N_ = 1 THEN DO ;
F = SQRT(1-RHO*RHO) ;
DO OVER XX ; MX = F * XX ; END ;
RHO = RHO * RES / F ;
END;
PROC PRINT ;
PROC REG ; MODEL MX .. RTX_M RTY_M KYL_M RES / NOINT ;