Implications of Survey Designs for Estimating Seasonal ARIMA Models

David A. Binder and J. Peter Dick
Statistics Canada

1. INTRODUCTION

It is common practice to analyze data from surveys where similar data items are collected on repeated occasions, using time series analysis methods. Most standard methods for these analyses assume the data are either observed without error or have independent measurement errors. However, in the analysis of repeated survey data, when there are overlapping sampling units between occasions, the survey errors can be correlated over time.

A commonly used model in the analysis of time series is the seasonal integrated autoregressive-moving average (ARIMA) regression model, which we discuss in this paper. We show how to incorporate the (possibly correlated) survey errors into the analysis. In particular, we consider the case where the survey (design) error can be assumed to be an ARMA process up to a multiplicative constant.

When such a model for the behaviour of the population characteristics is assumed, the minimum mean squared error, or, equivalently, the Bayes linear estimator for the characteristic at a point in time can be derived. This estimator incorporates the model structure which the classical estimators, such as the minimum variance linear unbiased estimators, ignore. When the model parameters are estimated from the survey data, the estimators are empirical Bayes.

Blight and Scott (1973), Scott and Smith (1974), Scott, Smith and Jones (1977), Jones (1980) and others considered the implications of certain stochastic models for the population means over time. In Binder and Dick (1989), these results were generalized using state space models and Kalman filters. In this paper, we extend the framework to include the model where differencing of the original series of the population means yields an ARMA model. We use the modified Kalman filter approach given by Kohn and Ansley (1986). To estimate the unknown parameters, we maximize the marginal likelihood function using the method of scoring. This approach can also handle missing data routinely. We also show how the survey estimates can be smoothed to incorporate the model features using empirical Bayes methods. Confidence intervals for these smoothed values are also given, using the method described by Ansley and Kohn (1986).

An example of this model is described in Section 5 using data from then Canadian Labour Force Survey. This example shows the implications on the estimates of the model parameters when the survey errors are taken into account. We also derive a smoothed estimate of the underlying process under the model assumptions.

2. THE MODEL

Suppose we have a series of point estimates from a repeated survey of a population characteristic, given by $y_1, y_2, \ldots, y_T$. We assume that y_t can be decomposed into three components, so that

$$y_t = x'_t\gamma + \theta_t + e_t, \qquad (2.1)$$

where $x'_t\gamma$ is a deterministic regression term, θ_t is a population parameter following a time series model, and e_t is the survey error, assumed to have zero expectation.

We first describe an integrated seasonal autoregressive-moving average model for $\{\theta_t\}$. We let B be the backshift operator; $\nabla = 1-B$ and $\nabla_s = 1-B^s$, where s is the seasonal period. We define the following polynomial functions:

$$\lambda(A) = 1 - \lambda_1 A - \lambda_2 A^2 - \ldots - \lambda_p A^P,$$

$$\alpha(A) = 1 - \alpha_1 A - \alpha_2 A^2 - \ldots - \alpha_p A^p,$$

$$\nu(A) = 1 - \nu_1 A - \nu_2 A^2 - \ldots - \nu_Q A^Q,$$

and

$$\beta(A) = 1 - \beta_1 A - \beta_2 A^2 - \ldots - \beta_q A^q.$$

The seasonal ARIMA $(p,d,q)(P,D,Q)_s$ model for $\{\theta_t\}$ is given by

$$\lambda(B^s)\alpha(B)\nabla^d\nabla_s^D\theta_t = \nu(B^s)\beta(B)\varepsilon_t, \qquad (2.2)$$

where the ε_t's are independent $N(0,\sigma^2)$. We define

$a(B) = \lambda(B^s)\alpha(B)$, a $(p+sP)$-degree polynomial;

$\Delta(B) = \nabla^d\nabla_s^D$, a $(d+sD)$-degree polynomial;

$b(B) = \nu(B^s)\beta(B)$, a $(q+sQ)$-degree polynomial;

$A(B) = a(B)\Delta(B)$, a $(p+d+sP+sD)$-degree polynomial;

$u_t = \Delta(B)\theta_t$, an ARMA$(p+sP,q+sQ)$ process.

Therefore, alternative representations of (2.2) are

$$a(B)\Delta(B)\theta_t = b(B)\varepsilon_t, \qquad (2.3)$$

$$A(B)\theta_t = b(B)\varepsilon_t, \qquad (2.4)$$

and

$$a(B)u_t = b(B)\varepsilon_t, \qquad (2.5)$$

We now consider the survey errors $\{e_t\}$ of expression (2.1). It will be assumed that the sample sizes of the repeated survey are sufficiently large that the errors for the survey estimates can be approximated by a multivariate normal distribution. In the simplest case, where the surveys are non-overlapping and the sampling fractions are small, the e_t's can be assumed to be independent. In a rotating panel survey, the survey errors are usually correlated. In this case, since the correlations between survey occasions are zero after panels have been rotated out, a pure moving average process can be used to describe the survey error process.

Alternatively, if a random sample of units are replaced on each survey occasion, a pure autoregressive process may best describe the process. More complicated models are also possible. For example, in a two-stage design, some of the first stage units may be replaced randomly on each occasion and the second stage units may have a rotating panel design. This might be represented by an autoregressive-moving average process.

In this paper, we assume that the survey error process is given by

$$e_t = k_t\omega_t, \tag{2.6}$$

where $\{\omega_t\}$ is an ARMA (m,n) process, given by

$$\phi(B)\omega_t = \psi(B)\eta_t \tag{2.7}$$

and

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_m B^m,$$

and

$$\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - \dots - \psi_n B^n,$$

The η_t's are independent $N(0,\tau^2)$. The factor k_t has been included in (2.6) to allow for non-homogeneous variances, even when the autocorrelation function is homogeneous in time.

In the model just described we assume that τ^2, the k_t's and the coefficients of $\phi(B)$ and of $\psi(B)$ can be estimated directly from the survey data, using design-based methods. However, in general, the other parameters are unknown. This includes γ, σ^2, and the coefficients of $\lambda(A)$, $\alpha(A)$, $\nu(A)$ and of $\beta(A)$. The x_t's is the regression term are assumed known.

3. STATE SPACE FORMULATION OF THE MODEL

3.1 General Formulation

The model described in Section 2 can be formulated as a state space model with partially improper priors. This has a number of advantages. It permits, through use of a modified Kalman filter, calculation of a marginal likelihood function, which can be maximized to estimate unknown parameters. It also accommodates smoothing of the original survey estimates, by removing the estimates of survey error from the data.

In the state space model, two processes occur simultaneously. The first process, the observation system, details how the observations depend on the current state of the process parameters. The second process, the transition system, details how the parameters evolve over time.

For the state space models we consider here, the observation equation is written as

$$y_t = \mathbf{h}_t' z_t \tag{3.1a}$$

and the transition equation is

$$z_t = F z_{t-1} + G\xi_t, \tag{3.1b}$$

where z_t is an $(r \times 1)$ state vector and $\mathbf{h}_t$ is a fixed $(r \times 1)$ vector. In the transition equation, $\mathbf{F}$ is a fixed $(r \times r)$ transition matrix, $\mathbf{G}$ is a fixed $(r \times m)$ matrix and the ξ_t's are independent normal vectors with mean zero and covariance $\mathbf{U}$.

The final requirement to complete the specification of the state space process is the initial conditions for z_0. In this paper, we shall use the improper prior formulation given in Kohn and Ansley (1986). In general, we assume that z_0 has a partially diffuse r-variate normal distribution with mean $\mathbf{m}(0|0) = 0$ and covariance matrix $\mathbf{V}(0|0)$, where

$$\mathbf{V}(0|0) = \kappa \mathbf{V}_1(0|0) + \mathbf{V}_0(0|0) \tag{3.2}$$

for large κ.

We denote the conditional mean of z_t given the observations up to and including time t' by $\mathbf{m}(t|t')$, and the conditional variance by $\mathbf{V(t|t')}$, where

$$\mathbf{V}(t|t') = \kappa\mathbf{V}_1(t|t') + \mathbf{V}_0(t|t'). \tag{3.3}$$

Recursive formulae for the cases where $t = t'$ and $t = t'+1$ are given in Kohn and Ansley (1986). They refer to this as the modified Kalman filter.

Since the model for $\{y_t\}$ given by (2.1) contains survey errors $\{e_t\}$ an estimate of the components without survey error, given by

$$y_t \text{ (smoothed)} = \mathbf{x}_t'\gamma + \theta_t \tag{3.4}$$

is often of interest. When the right hand side of (3.4) can be expressed as $\mathbf{g}_t' z_t$, for some $\mathbf{g}_t'$, then it is possible to obtain the conditional mean and variance of the linear combination $\mathbf{g}_t' z_t$ given all the data, using the modified Kalman filter. To do this, the recursions are applied up to time t to obtain $\mathbf{m}(t|t)$ and $\mathbf{V}(t|t)$. Then the state vector z_t is augmented by the state $z_{t,r+1} = \mathbf{g}_t' z_t$, and $\mathbf{m}(t|t)$ and $\mathbf{V}(t|t)$ are also appropriately augmented. The matrix $\mathbf{F}$ in (3.1b) is modified to add the equation $z_{t+1,\ r+1} = z_{t,r+1}$. After these modifications, the modified Kalman filter can be used as before, so that the last component of $\mathbf{m}(T|T)$ gives the conditional expectation of $\mathbf{g}_t' z_t$, given all the data, $y_1, y_2, \dots y_T$. As well, the last diagonal component of $\mathbf{V}(t|t)$ gives the conditional variance. This procedure can be generalized to include any number of smoothed estimates and their conditional covariances.

3.2 Model for θ

Harvey and Phillips (1979) described a method to put the ARIMA model (2.4) into the state space form given by (3.1). The dimension of z_t is $r = \max(p+d+sP+sD, q+sQ)$. By augmenting $\mathbf{A} = (A_1, \dots, A_{p+d+sP+sD})$ or $\mathbf{b} = (b_1, \dots, b_{q+sQ})$ with zeroes to have dimension r, the ARIMA model may be written in the form given by (3.1), where $\mathbf{h}_t' = (1, 0, \dots, 0)$, $\mathbf{G}_t' = (1, -b_1, \dots, -b_{r-1})$ and

$$F = \left[\begin{array}{c|c} A_1 & \\ \vdots & I_{r-1} \\ A_{r-1} & \\ \hline A_r & 0' \end{array}\right],$$

where I_{r-1} is the $(r-1) \times (r-1)$ identity matrix and $0'$ is a row vector of zeroes.

In this formulation, the state vector $z_t = (z_{1t}, \dots, z_{rt})'$ is defined as

$$z_{it} = A_i\theta_{t-1} + A_{i+1}\theta_{t-2} + \dots + A_r\theta_{t-(r-i+1)}$$

$$- b_{i-1}\varepsilon_t - b_i\varepsilon_{t-1} - \dots - b_{r-1}\varepsilon_{t-(r-i)}, \tag{3.5}$$

for $i = 2, 3, \dots, r$ and $z_{1t} = \theta_t$.

To complete the specification for $\{\theta_t\}$, initial conditions for z_0 are required. These are given in Ansley and Kohn (1985), a summary of which is provided here.

From expression (2.5), $\{u_t\}$ is an ARMA process. We define

$$\theta\_ = (\theta_0, \theta_{-1}, \ldots, \theta_{-S})',$$

where $S = \max(0, p+sP+d+sD-1)$. We let

$$u\_ = (u_0, u_{-1}, \ldots, u_{-R})',$$

where $R = \max(0, p+sP-1)$. Finally, we let

$$w\_ = (\theta_{-R-1}, \theta_{-R-2}, \ldots, \theta_{-S})',$$

when $S > R$.

Now, $u\_$ is assumed to be a stationary ARMA process, so that its covariance matrix can be derived from expression (2.5). It is assumed that $w\_$ is $N(0, \kappa I)$ and is independent of $u\_$. Since $(u_', w_')'$is a linear combination of $\theta\_$, the covariance matrix for $\theta\_$ can be derived. Using the form of expression (3.5) for z_0, the initial covariance matrix can be computed. Note that when both d and D are zero, so that no differencing takes place in the model, then $w\_$ is the null vector and we have $u\_ = h\_$.

3.3 Model for the Observed Data

In Section 2 we assumed that $e_t = k_t \omega_t$, where ω_t is an ARMA(m,n) model. Therefore, from the discussion in Section 3.3, it is clear that e_t can be represented in state space form, with $h_t = (k_t, 0, \ldots, 0)'$, and $e_t = h_t' z_t$.

The regression component can be similarly represented. We let $z_0 = \gamma$, the regression coefficients, assumed to have mean zero and covariance κI. The transition equation is simply $z_{t+1} = z_t$.

Since we can represent each of the components of y_t in expression (2.1) by a state space model, it straightforward to combine the individual models into an overall model, by extending the state vector to include the state vectors from the individual components. The observation equation is then the sum of the three individual components.

4. ESTIMATION OF THE STATE SPACE MODEL

4.1 Estimation of the Parameters

The unknown parameters of this model are σ^2, and the coefficients of $\lambda(A)$, $\alpha(A)$, $\nu(A)$and $\beta(A)$. We performed the iterations on $\log(\sigma^2)$, rather than σ^2, to avoid problems with negative values. Note that the regression coefficients, γ, are included as parameters of the state vector. The model for the vector of observations $y = (y_1, y_2, \ldots, y_T)'$ given in Section 3 is equivalent to

$$y = M\eta + \zeta, \tag{4.1}$$

where η is j-variate $N(0, \kappa I)$, ζ is T-variate $N(0, W)$, and M is a $T \times j$ matrix.

Kohn and Ansley (1986) recommended maximizing the limit of $\kappa^{j/2}$ times the likelihood function for the data, as κ tends to infinity. It can be shown that the limit of the likelihood function is equivalent to the marginal likelihood function of $y - M\hat{\eta}$, where $\hat{\eta}$ is the maximum likelihood estimate of η when M and W are known. Tunnicliffe-Wilson (1989) has shown that the Jacobian of transformation from the data y to $(\hat{\eta}, y - M\hat{\eta})$ does not depend on the model parameters of W whenever M is known. As well, the derivative of the transformation from y to $\hat{\eta}$ is M. Ansley and Kohn (1985) has shown that M does not depend on the unknown parameters. By using the modified Kalman filter, the computations for the marginal likelihood function are straightforward.

The procedure we employed computes both the marginal likelihood function and its first derivatives with respect to the unknown parameters. This involves taking first derivatives of the initial conditions and of $m(t|t')$ and the components of $V(t|t')$ for $t=t'$ and $t=t'+1$. All the computations were done using PROC IML in SAS.

The likelihood function was maximized using a modification of the method of scoring. This modification allowed for varying step sizes. On each iteration, the likelihood function was computed at the previous step size, as well as at this step size multiplied and divided by a predetermined constant. (We used 1.1 as the factor.) The next step size was that which maximized the likelihood function among the three points. Each time a check was made to determine whether the parameters were in range. This was done by checking for positive semi-definiteness of the initial covariance matrix of the state vector. If it was out of range, the step size was divided again by the constant and the procedure repeated.

To obtain the estimated variance matrix for the estimated parameters, the inverse of the Fisher information was used. This is readily computed since the first derivatives of the likelihood function are available.

4.2 Estimation of the Smoothed Values

Smoothed values for the estimates can be obtained by zeroing out that component of the state vector which corresponds to the survey error. However, this still leaves open the question of how to estimate its variance. To derive the standard error of the smoothed estimate it is necessary to account for the fact that the unknown parameters have been estimated from the data, particularly when the data series is short; see Jones (1979).

To obtain the variance of $g'z_t$, it is sufficient to derive the variance $z_T - \hat{m}(T|T)$, where $\hat{m}(T|T)$ is the estimate of $m(T|T)$ at the estimated parameter values. This is because the state vector has been augmented to include $g'z_t$. Now,

$$z_T - \hat{m}(T|T) = [z_T - m(T|T)] + [m(T|T) - \hat{m}(T|T)]. \tag{4.2}$$

The first component of the right hand side of (4.2) has conditional variance $\mathbf{V}(T|T) = \mathbf{V}_0(T|T)$, assuming that $\mathbf{V}_1(T|T) = 0$. The second component of (4.2) represents a bias term and is independent of the first term, since it depends only on the data y. By taking a Taylor series expansion of the second term around the true parameter values and ignoring higher terms, we have the second component of (4.2) is

$$\mathbf{m}(T|T) - \hat{\mathbf{m}}(T|T) = [\frac{-\partial\hat{\mathbf{m}}(T|T)}{\partial\alpha}]' (\hat{\phi} - \phi), \quad (4.3)$$

where ϕ is the vector of unknown parameters and $\hat{\phi}$ is its estimate. Therefore, the variance of (4.2) is approximately

$$\mathrm{Var}[z_T - \hat{\mathbf{m}}(T|T)] = \mathbf{V}_0(T|T) + [\frac{\partial\hat{\mathbf{m}}(T|T)}{\partial\phi}]' \mathbf{V}_\phi [\frac{\partial\hat{\mathbf{m}}(T|T)}{\partial\phi}] \quad (4.4)$$

where $\mathbf{V}_\phi$ is the covariance matrix for the unknown parameters. Expression (4.4) is estimated by using the estimated parameter values. This is the same approach as that given by Ansley and Kohn (1986).

5. LABOUR FORCE SURVEY DATA

To demonstrate this procedure, we took data from the Canadian Labour Force Survey (LFS). The LFS is a monthly rotating panel survey. Each panel, which contains one-sixth of the selected households remain in the sample for six consecutive months. The sample design is a stratified multi-stage design. The primary sampling units are rotated out after approximately two years.

The data were from the ten years from January 1977 to December 1986. We used the series of number of employed for the province of Nova Scotia and from the subprovincial area within Nova Scotia corresponding to Cape Breton Island. This province was chosen because the sampling errors were moderate compared to the larger provinces and because subprovincial data were available. The models were fitted to the logarithm of the series.

Lee (1987) estimated the autocorrelations for Nova Scotia up to a lag of eleven. Using these autocorrelations, we estimated the coefficients of $\phi(B)$ and of $\psi(B)$ given in (2.7) and we estimated τ^2. A good fit was found using an ARMA(1,6) model. The estimated parameters were $\phi_1 = 0.7322$, $\psi_1 = -0.005589$, $\psi_2 = -0.02736$, $\psi_3 = -0.06153$, $\psi_4 = -0.03175$, $\psi_5 = -0.03184$, $\psi_6 = -0.06027$, and $\tau^2 = 0.4160$. The k_t's of (2.6) were the estimated standard errors of the estimates, taking a Taylor series approximation for the logarithms.

A series of models were fitted to the data where no sampling error was assumed; that is, all the k_t's were taken as zero. These models were then refitted using the assumed structure for the survey error. We compared the estimated parameter values. As well in the case where the survey error structure is assumed to be non-zero, we computed smoothed values for the survey estimates and compared their standard errors with the standard errors of the original series.

After some model fitting, ignoring the survey error component, a model selected for the Nova Scotia series was a seasonal ARIMA $(2,1,0)(0,1,1)_{12}$. However, the seasonal moving average component converged to 1, implying that a deterministic regression term rather than differencing should be used to account for the seasonality. A seasonal ARIMA $(2,1,0)(0,0,1)$ with a regression term was fitted. The 12 regression variables included a linear trend and a dummy variable for each of the the first 11 months. The dummy variable for a reference month took the value 1, -1 or 0, for the reference month, for December and for the other months, respectively. (Note that an intercept term is not estimable because the first differences for the data are fitted.) The same model was then fitted to the Cape Breton Island data.

The parameter estimates for both Nova Scotia and Cape Breton Island are displayed in Table 1. We display the estimates which do not take into account the survey error component in the "Without Sampling Error" columns.

The estimated model, when the sampling errors are taken in account, is strikingly different. In both series, the estimates for the ARIMA parameters are all zero, implying that there is no ARIMA component. (Note that when the model variance is zero, the other ARIMA model parameters are no longer identifiable.) The regression parameter estimates are similar to the estimates obtained by ignoring the sampling error component. This is because the estimates are unbiased under either model assumption. However, the t-values for the regression coefficients are too large when the survey error component is ignored.

In summary, when the sampling errors component is incorporated, the best model will differ from the case where the sampling errors are ignored. Instead of including an ARIMA term, the fitted model contains only a deterministic regression component along with the sampling error component. In effect, the component from the ARIMA model which is found when the sampling error is ignored is small compared to the survey error in the data.

Once the parameters are estimated, the smoothing procedure described in Section 4.2 was applied to the two series. The variance reduction of the smoothed values was substantial, ranging from an 80% to 95% reduction over the original survey error variances. Of course, this reduction makes strong assumptions about the validity of the model, which could easily be violated. In fact, the fitted model, consisting of only a deterministic regression term, seems unrealistic. However, for analytical purposes it is quite revealing.

REFERENCES

Ansley, C.F. and R. Kohn. 1985. A structured state space approach to computing the likelihood of an ARIMA process and its derivatives. Journal of Statistical Computation and Simulation. 21: 135-169.

Ansley, C.F. and R. Kohn. 1986. Prediction mean squared error for state space models with estimated parameters. Biometrika. 73: 467-473.

Binder, D.A. and J.P. Dick. 1989. Modelling and estimation for repeated surveys. Survey Methodology. 14: to appear.

Blight, B.J.N. and A.J. Scott. 1973. A stochastic model for repeated surveys. Journal of the Royal Statistical Society, Series B. 35: 61-68.

Harvey, A.C. and G.D.A. Phillips. 1979. Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. Biometrika. 66: 49-58.

Jones, R.G. 1979. The efficiency of time series estimators for repeated surveys. Australian Journal of Statistics. 21: 45-56.

Jones, R.G. 1980. Best linear unbiased estimators for repeated surveys. Journal of the Royal Statistical Society, Series B. 42: 221-226.

Kohn, R. and C.F. Ansley. 1986. Estimation, prediction and interpolation for ARIMA models with missing data. Journal of the American Statistical Association. 81: 751-761.

Lee, H. 1987. Estimation of panel correlations for the Canadian Labour Force Survey. Technical Report. Statistics Canada.

Scott, A.J. and T.M.F. Smith. 1974. Analysis of repeated surveys using time series methods. Journal of the American Statistical Association. 69:674-678.

Scott, A.J., T.M.F. Smith and R.G. Jones. 1977. The application of time series methods to the analysis of repeated surveys. International Statistics Review. 45: 13-28.

Tunnicliffe-Wilson, G. 1989. On the use of marginal likelihood in time series model estimation. Journal of the Royal Statistical Society, Series B. 51: 15-27.

TABLE 1. Parameter Estimates

	Nova Scotia				Cape Breton Island			
	Without Sampling Error		With Sampling Error		Without Sampling Error		With Sampling Error	
Parameter	Estimate	T-value	Estimate	T-value	Estimate	T-value	Estimate	T-value
Alpha (1)	-0.19	-1.9	-	-	-0.17	-1.8	-	-
Alpha (2)	0.01	0.1	-	-	-0.08	-0.8	-	-
Nu	0.08	0.7	-	-	-0.12	-1.1	-	-
Sigma	0.009	-	0.000	-	0.031	-	0.000	-
Trend	0.0013	2.0	0.0011	3.7	0.0017	0.7	0.0008	1.4
January	-0.06	-22.2	-0.06	-5.9	-0.07	-7.5	-0.07	-3.5
February	-0.06	-21.8	-0.06	-5.5	-0.08	-8.2	-0.09	-4.2
March	-0.05	-20.8	-0.05	-5.3	-0.08	-8.1	-0.09	-3.8
April	-0.04	-15.2	-0.04	-4.0	-0.05	-5.6	-0.06	-3.4
May	0.01	3.7	0.01	1.0	0.02	2.3	0.02	1.0
June	0.04	15.8	0.04	4.4	0.06	6.4	0.06	2.8
July	0.07	26.0	0.07	7.6	0.11	11.0	0.10	5.0
August	0.07	27.0	0.07	8.1	0.10	10.7	0.10	5.3
September	0.02	10.0	0.03	3.1	0.03	3.3	0.03	1.2
October	0.02	6.0	0.02	1.8	0.01	1.0	0.01	0.6
November	-0.003	-1.2	-0.004	-0.4	-0.01	-1.5	-0.001	-0.1

A KALMAN FILTER APPROACH TO LABOR FORCE ESTIMATION USING SURVEY DATA

Richard Tiller, Bureau of Labor Statistics
441 G Street NW, Washington, DC 20212

Abstract

A new time series method for estimating employment and unemployment in 40 States was introduced by the Bureau of Labor Statistics in 1989. It uses the Kalman filter to combine current period State-wide estimates from the Current Population Survey with past sample estimates and auxiliary data from the unemployment insurance system and the Current Employment Statistics payroll survey. The purpose is to reduce high variance in the CPS labor force estimates due to small sample sizes. This paper discusses the basic time series approach used and presents the unemployment model as an example.

KEY WORDS: Time series, correlated measurement error, state space models

1.0 Introduction

In January 1989, the Bureau of Labor Statistics (BLS) introduced a new method for estimating monthly employment and unemployment for 39 States and the District of Columbia. The new method uses time series models fitted to the statewide monthly sample data from the Current Population Survey (CPS). The purpose of this paper is to provide information on the basic modeling approach used and on the current and planned research to develop further improvements. The unemployment rate models are presented as examples.

The most direct way to estimate the characteristics of a population, such as labor force status, is to conduct a large-scale sample survey based on a probability design. Often times reliable estimates are available for a large area but the sample is too thinly spread to provide reliable estimates for subareas. For periodic surveys, time series techniques have received increasing interest as a way of making extensive use of whatever data are available from the survey specific to subareas. The CPS provides an example of a periodic survey that is particularly well-suited to the application of these techniques. Each month, a sample of about 59,000 households is interviewed to provide estimates of the labor force status of the population. Reliable monthly estimates are produced for the nation as a whole and for eleven of the more populous States. For the remaining 40 States (including the District of Columbia), the sample is not large enough to support direct use of the monthly estimates.

Prior to 1989, labor force estimates for the 40 States were based on the Handbook method (Bureau of Labor Statistics, 1988). This method used as its primary inputs data on a count of workers drawing unemployment insurance (UI) benefits and estimates of nonagricultural payroll employment from the Current Employment Statistics (CES) survey.

The new approach to estimation is based on a signal plus noise model that treats the monthly CPS sample data as the sum of a stochastically varying true labor force series (signal) and error (noise) generated by the CPS sampling process. Monthly CPS labor force estimates along with sample design information are combined with UI and CES data in a time series model of the data generating process. The basic idea is to reduce the effects of high variance in the CPS due to small sample sizes by using both current and past sample data along with auxiliary data in a more systematic way than was done before. Given a model describing the dynamic behavior of the unobserved population series and autocovariances of the sample error, the Kalman filter (KF) may be used to estimate the true series. The KF has a number of particularly useful features: It allows for a wide variety of approaches to the specification of the signal and noise components; its recursive structure provides a very efficient algorithm for the preparation of labor force estimates each month by 40 State agencies; and finally, the KF is a very useful tool for implementing estimators of the unknown parameters of dynamic models.

The remainder of this paper is organized in the following way: Section 2 presents the basic signal plus noise model in a state space framework; section 3 discusses practical implementation issues; section 4 presents an application of the model to estimating unemployment; and finally section 5 discusses current and future research plans.

2.0 Time Series Approach to Modeling CPS Data

The probability designed CPS yields monthly estimates of the labor force characteristics of each State's population. The classical survey sample approach treats the true labor force values as fixed and focuses on the variation due to sampling. The time series approach, as exemplified by Scott and Smith (1974) and Bell and Hillmer (1987b), treats the unobserved values estimated by sample surveys as varying stochastically over time. From this perspective, the data generating process giving rise to a State's CPS labor force series consists of a stochastically varying true labor force (signal) and measurement error (noise) generated by the CPS survey design. The time series approach seeks to synthesize two different approaches to estimation by using time series theory to model the signal component and information from the sample survey to specify the noise component of the observed sample series.

2.1 Signal Component of the CPS

A dynamic linear regression approach is used to model the true values of the employment level and the unemployment rate for each of the 40 States. Since each is estimated using a model of the same general form, we will first discuss those features common to both models and then use the unemployment rate as an example.

The observed CPS labor force estimate, Y_t, is represented as the sum of the signal, θ_t, plus a noise term, e_t,

$Y_t = \theta_t + e_t$.

The signal, or true labor force is specified as generated by a dynamic linear model consisting of a time varying mean $\mu_{t/X}$ and a disturbance u_t,

$$\theta_t = \mu_{t/X} + u_t \tag{1}$$

The mean represents that part of θ_t that can be "explained" by the observed X variables,

$\mu_{t/X} = X_t \beta_t$

where,

X_t = 1 x k vector of observed regressor variables

β_t = k x 1 vector of stochastic coefficients.

The presence of these variables serves two important and related functions. First, it allows the use of auxiliary data obtained through administrative and other non-CPS sources to improve the efficiency of model estimates. Secondly, as economic indicators, these variables play a useful descriptive function that helps State analysts explain their labor force movements. (The specific variables used as regressors will be discussed later for the unemployment rate model.)

The regression coefficients are treated as varying stochastically according to a first order vector autoregressive process (VAR),

$$\beta_t = T_\beta \beta_{t-1} + v_{\beta t} \tag{2}$$

where,

T_β = k x k matrix of fixed parameters