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

$v_{\beta t}$ = k x 1 vector of white noise coefficient disturbances.

The basic role of this variation is to represent uncertainty in the ability of the model to fully depict the true labor force. For example, changes in a coefficient may originate from omitting a variable whose correlations with the regressors change over the sample period (Engle and Watson, 1985). As discussed below, increasing these variances will tend to discount the model based estimates, thereby weighting the individual current period CPS sample estimates more heavily.

The disturbance, u_t, in (1) is the "errors in equation," or the stochastic part of the signal that is not accounted for by either the X variables or coefficient variation. Since this disturbance may be autocorrelated, it is represented as a general ARMA (p_u, q_u) process,

$$u_t = \phi_u(L)\,\Psi^{-1}(L)\,v_{ut} \tag{3}$$

where,

u_t = errors in equation

v_{ut} = white noise disturbance to u_t

$\Psi_u(L) = 1 - \sum_{i=1}^{p_u} \Psi_{ui} L^i$, autoregressive operator for u_t

$\phi_u(L) = 1 + \sum_{i=1}^{q_u} \phi_{ui} L^i$, MA operator for u_t

L = lag operator, such that $L^i Y_t = Y_{t-i}$.

The random disturbances, $v_{\beta t}$ and v_{ut}, are assumed to have zero means and be mutually independent,

$$\begin{bmatrix} v_{\beta t} \\ v_{ut} \end{bmatrix} \sim ID \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & \sigma_{v_u v_u} \end{bmatrix} \right\} \tag{4}$$

where,

$Q = Cov(v_{\beta t}) = Diag(\sigma_{\beta_1\beta_1}, \cdots, \sigma_{\beta_k\beta_k})$.

2.2 Noise Component of the CPS

The role of the noise component is to incorporate important features of the CPS sample design and estimation procedures directly into the modeling process. The noise component is treated as due entirely to error arising from sampling only a portion of the entire population. While other sources of error may be important, we do not deal with them here. For our purposes, we view the "true" labor force value in a practical context as that which would be measured by expanding the CPS sample to include the entire population.

The CPS is a complex multi–stage sample of the population; in the first stage a stratified sample of PSU's are selected; and within PSU's housing units are selected from a stratified systematic cluster sample with only a partial replacement of housing units each month. The CPS estimation process consists of a noninterview adjustment, two stages of ratio adjustments, and a compositing procedure that takes into account the sample overlap in adjacent months (Bureau of the Census, 1978). Given that Y_t is the CPS sample estimate of the total number of persons in the population having a specific labor force characteristic, θ_t, the sample error,

$e_t = Y_t - \theta_t$

has variance,

$$\sigma_{ee,t} = D_y S_y^2 \tag{5}$$

and covariance function,

$\gamma_{ts} = Cov(e_t, e_s)$ for any t,s

where,

D_y = ratio of the variance of the CPS estimator to the variance of the simple random sample (SRS) estimator (design effect)

$S_y^2 = (N_t / n_t)\,\theta_t\,(1 - P_t)$

N_t = population size

n_t = sample size

$P_t = \theta_t / N_t$.

As the above equations illustrate, the CPS sample error has both a heteroscedastic and autocorrelated structure. Equation (5) illustrates three major sources of heteroscedasticity: (1) sample redesigns as reflected by changes in D_y; (2) changes in the sample interval N_t / n_t; and (3) changes in the true values θ_t and P_t. The first two cause discrete shifts in the sample variance. For example, the CPS is redesigned each decade to make use of decennial census data to update the sampling frame and estimation procedures. Most recently, a State based design was phased in during 1984/85 along with improved procedures for noninterviews, ratio adjustments and compositing. Changes in State sample sizes have occurred more frequently than redesigns and have had a major impact on variances at the State level. Even with a fixed design and sample size, the error variance will be changing because it is a function of the size of the true labor force. Since the labor force is both highly cyclical and seasonal, we can expect the variance to follow a similar pattern.

The autocovariance structure of e_t is influenced primarily by three things. First, the monthly sample is composed of 8 independent subsamples of housing units known as rotation groups. A rotation group is interviewed for 4 months, dropped from the sample for 8 months and then returned for 4 months. Clearly, correlations will be generated since identical housing units will appear in more than one monthly sample. The 4–8–4 feature of the rotation scheme will produce correlations over a 15 month period with the largest coming at 1 month (75% overlap) and at 2 and 12 months (50% overlap).

Secondly, the use of a rotation system requires the periodic selection of additional samples. When a cluster of housing units permanently drops out of a rotation group it is replaced with nearby units. Since the new units will have characteristics similar to those being replaced this will result in correlations between non–identical households in the same rotation group (Train, Cahoon and Makens, 1978).

Finally, the dynamics of the sample error will also be affected by the composite estimator. This is a weighted average of an estimate based on the entire sample for the current month only and an estimate which is a sum of the prior month composite and change that occurred in the 6 rotation groups common to both months (Bureau of the Census, 1978).

Both the heteroscedastic and autoregressive behavior of the CPS estimator may be accounted for by modeling e_t in multiplicative form (Bell and Hillmer, 1989),

$e_t = \gamma_t e_t^*$

with e_t^* following an ARMA process with constant variance,

$$e_t^* = \phi_e(L)\,\Psi_e^{-1}(L)\,v_{et} \tag{6}$$

$v_{et} \sim NID(0, \sigma_{v_e v_e})$

$\sigma_{e^*e^*} = \sigma_{v_e v_e} \sum_{k=0}^{\infty} g_k^2$

where the weights $\{g_k\}$ are computed from the generating function,

$g(L) = \phi_e(L)\,\Psi_e^{-1}(L)$.

The heteroscedastic component of e_t is the square root of the ratio of variances,

$$\gamma(t) = \sqrt{\frac{\sigma_{ee,t}}{\sigma_{e^*e^*}}}.$$

The autocorrelation structure of e_t is also likely to be affected by changes in sample design. For example, in the most recent redesign, the weights in the composite estimator were altered. As suggested by Bell and Hillmer (1989), one way to control for this is to treat the ARMA coefficients as constants that shift with major redesigns but are stable for a given design.

2.3 State Space Form and the Kalman Filter Algorithm

The signal and noise components of the labor force model may be put into state space form. We will first consider a very general form that, while not suitable for estimation, is useful for demonstrating its flexibility and discussing the kinds of restrictions that must be imposed for practical implementation.

In a state space formulation, the unobserved signal and noise are the state variables whose evolution over time is described by a set of transition equations. An observation equation transforms the state variables into the observed sample series.

The transition equations in a state space system must take the form of a first order VAR. For our problem, the unobservable variables are β_t, u_t and e^*_t. The coefficient vector, β_t, is already in the appropriate form as can be seen from (2). While u_t and e^*_t have been specified as ARMA processes in (3) and (6), they may be converted into vectors, S_{ut} and S_{et} respectively, that follow a first order VAR form. The basic rule is that any ARMA (p, q) process can be converted to an r x 1 first order VAR, r = max (p, q + 1), see Harvey (1981). The transition equations are given below, where S_t is the state vector consisting of β_t, S_{ut} and S_{et}.

$$\underset{(m\times 1)}{S_t} = \underset{(m\times m)}{T_t}\; S_{t-1} + \underset{(m\times \ell)}{\Gamma_t}\; \underset{(\ell\times 1)}{v_t}$$

$$\begin{bmatrix} \beta_t \\ S_{ut} \\ S_{et} \end{bmatrix} = \begin{bmatrix} T_\beta & & 0 \\ & T_u & \\ 0 & & T_{et} \end{bmatrix} \begin{bmatrix} \beta_{t-1} \\ S_{ut-1} \\ S_{et-1} \end{bmatrix} + \begin{bmatrix} I_k & & 0 \\ & \Gamma_u & \\ 0 & & \Gamma_{et} \end{bmatrix} \begin{bmatrix} v_{\beta_t} \\ \hline v_{ut} \\ v_{et} \end{bmatrix}$$

$E(v_t v_t') =$ block diagonal $(Q, \sigma_{v_u v_u}, \sigma_{v_e v_e, t})$

where,

$$T_u = \left[\begin{array}{c|c} \Psi_{u1} & I_{r_u-1} \\ \vdots & \\ \hline \Psi_{ur_u} & 0 \end{array}\right], \quad T_{et} = \left[\begin{array}{c|c} \Psi_{e1,t} & I_{r_e-1} \\ \vdots & \\ \hline \Psi_{er_e,t} & 0 \end{array}\right],$$

$$\Gamma_u = \begin{bmatrix} 1 \\ \hline \phi_{u1} \\ \vdots \\ \phi_{ur_u-1} \end{bmatrix}, \quad \Gamma_{et} = \begin{bmatrix} 1 \\ \hline \phi_{e1,t} \\ \vdots \\ \phi_{er_e-1,t} \end{bmatrix}$$

$m = k + r_u + r_e$

k = number of regressor variables

$\ell = k + 2$

$r_u = \max(p_u, q_u + 1)$, p_u, q_u are order parameters of the ARMA form of of u_t

$r_e = \max(p_e, q_e + 1)$, p_e, q_e are order parameters of the ARMA form of e_t

The observation equation, via the selection vector H_t, takes linear combinations of the state variables to form the signal and noise components which sum to the observed series.

$Y_t = H_t S_t = \theta_t + e_t$

$\theta_t = H_{\theta t} S_t$

$e_t = H_{et} S_t$

where,

$$\underset{(1\times m)}{H_t} = \left[X_t \mid 1 \mid 0_{r_u-1} \mid \gamma_t \mid 0_{m-k-r_u-2}\right],$$

$$H_{\theta t} = \left[X_t \mid 1 \mid 0_{m-k-1}\right], \quad H_{et} = \left[0_{k+r_u} \mid \gamma_t \mid 0_{r_e-1}\right]$$

Given the state space form of the unobserved signal and noise components, the KF provides a means of estimating them by conditioning on one sample observation at a time. To specify this algorithm, let the expectation of S_t conditioned on the observed data up to time t – j be given by

$S_{t/t-j} = E(S_t \mid Y_{t-j}, \cdots, Y_1)$

with covariance matrix,

$P_{t/t-j} = E\left\{(S_t - S_{t/t-j})(S_t - S_{t/t-j})' \mid Y_{t-j}, \cdots, Y_1)\right\}$

and a prediction of the sample estimate Y_t given its past values by

$Y_{t/t-1} = H_t S_{t/t-1}$

with variance,

$E(Y_t - Y_{t/t-1})^2 = H_t P_{t/t-1} H_t' = f_{t/t-1}$.

Given an estimate of S_t based on data up to but not including the t^{th} observation, an estimator of S_t involving current data is formed as a weighted average of $S_{t/t-1}$ and the current sample estimate Y_t,

$$S_{t/t} = (I - K_t H_t) S_{t/t-1} + K_t Y_t$$
$$= S_{t/t-1} + K_t (Y_t - Y_{t/t-1})$$

with covariance matrix,

$P_{t/t} = (I - K_t H_t') P_{t/t-1}$

where $S_{t/t-1}$ and $P_{t/t-1}$ are estimated recursively from

$S_{t/t-1} = T S_{t-1/t-1}$

$P_{t/t-1} = T P_{t-1/t-1} T' + \Gamma E(v_t v_t') \Gamma'$.

The weighting vector K_t, known as the gain of the filter, is given by

$K_t = P_{t/t-1} H_t' / f_{t/t-1}$

where the elements of K_t are determined by minimizing the sum of the diagonal elements of $P_{t/t}$ (Gelb, 1974).

Using the KF equations, the observed sample estimate at time t is separated into its signal and noise components,

$Y_t = \theta_{t/t} + e_{t/t}$

$\theta_{t/t} = \theta_{t/t-1} + h_{\theta t} \tilde{Y}_t$

$e_{t/t} = e_{t/t-1} + (1 - h_{\theta t}) \tilde{Y}_t$

where,

$\tilde{Y}_t = Y_t - Y_{t/t-1}$

$$h_{\theta t} = H_{\theta t} K_t = \left[\mathrm{Var}\,(\theta_t / \theta_{t-1}) + H_{\theta t}\, T\, P_{t-1/t-1}\, T'\, H_t \right] / f_{t/t-1}$$

$$1 - h_{\theta t} = H_{et} K_t = \left[\sigma_{ee,t} + H_{et}\, T\, P_{t-1/t-1}\, T'\, H_t \right] / f_{t/t-1}$$

$$\mathrm{Var}\,(\theta_t / \theta_{t-1}) = \sum_{i=1}^{k} X_{it}^2\, \sigma_{\beta_i \beta_i} + \sigma_{v_u v_u}$$

$$\sigma_{ee,t} = \gamma_t^2\, \sigma_{v_e v_e}.$$

The weight, $h_{\theta t}$, decomposes the prediction error $\tilde{Y}_t$ into its signal and noise components, respectively. This decomposition illustrates how the KF combines a time series estimator, $\theta_{t/t-1}$, with a current sample estimate, Y_t, to produce a minimum mean square error estimator of the signal. The amount by which $\theta_{t/t-1}$ is adjusted toward Y_t is a function of the size of the time series variance component, Var $(\theta_t / \theta_{t-1})$, relative to the heteroscedastic sample error variance, $\sigma_{ee,t}$. A relatively large value of $\sigma_{ee,t}$ results in a small value for $h_{\theta t}$ and hence only a small adjustment to the time series predictor $\theta_{t/t-1}$ in forming $\theta_{t/t}$. Conversely, if the sample variance is small, $\theta_{t/t}$ will not differ much from the current sample estimate Y_t.

The KF provides the minimum mean square error of the state vector, S_t, based on all sample data through time t, in a recursive manner, and is thus ideally suited to real time situations where new sample data become available each period. However, as data become available after time t, the estimate $S_{t/t}$ will not incorporate this new information, since the KF only moves forward in time. The suboptimality of previous period estimates is easily remedied through a process called smoothing.

The basic type of smoothing (fixed interval) relevant to this paper can be described in simple terms as combining two types of KF estimators (Maybeck, 1979). The first is the forward filter estimate, previously described, which at time t is based on all past and present sample data, $S_{t/t}$. The second is a backward filter, which is the KF run in reverse, starting at the end of the sample period, say t = n, and proceeding to the beginning, producing at each t predictions based on only future data (relative to the forward filter). Let the backward filter be denoted by $S_{t/t+1}$. The optimal smoothed estimator is then formed by combining the estimates from the two filters, in proportion to their mean square errors, as shown below.

$$S_{t/n} = P_{t/n} \left[P_{t/t}^{-1} S_{t/t} + (P_{t/t+1})^{-1} S_{t/t+1} \right]$$

$$P_{t/n} = \left[P_{t/t}^{-1} + (P_{t/t+1})^{-1} \right]^{-1}$$

From the covariance expression for $S_{t/n}$, we have $P_{t/n}^{-1} = P_{t/t}^{-1} + (P_{t/t+1})^{-1}$ which implies that $P_{t/n} - P_{t/t}$ is negative semidefinite. Thus, the smoothed estimator of S_t is never worse than the forward filter estimator. In fact, it is generally much better except for the last data point where the two estimators are identical. For this reason, historical labor force estimates are produced by the smoothing algorithm.

3.0 Practical Implementation Issues

The state space formulation allows for considerable flexibility in specifying the signal component. It includes as special cases two classes of model based approaches to sample surveys that have appeared in the literature.

If the variances of stochastic coefficient change are set to zero, i.e. Q = 0 and e_t and u_t are white noise, the system reduces to Ericksen's (1974) sample regression model. In this case, the signal extraction problem is solved by fitting a weighted least squares equation to the observed sample data.

By setting β_t and Q to zero, we have a class of models based on Wiener–Kolmogorov signal extraction theory. The regression mean drops out and the signal reduces to a covariance stationary process. If in addition, the variance and the ARMA parameters of the e_t process are also held constant, then e_t will also be covariance stationary. Scott and Smith (1974) adapted the classical signal extraction approach to survey data where the covariances have to be estimated. Bell and Hillmer (1987a) discuss ways of handling nonstationarity in the signal process.

In the model developed in this paper, nonstationarity in the signal is handled by the regressors and by stochastic changes in their coefficients. The transition equation (2) governing the behavior of the coefficients can accommodate a wide variety of patterns (Los, 1985). In practice, restrictions must be imposed to reduce this number to a manageable size. We specify these coefficients to follow independent random walks, i.e., $T_\beta = I$ and Q is diagonal. This specification has several practical advantages (Engle and Watson, 1985). It is parsimonious, involving only k parameters, change tends to be smooth from month to month, but over long periods the model is allowed to adapt to fundamental structural change.

If the state space parameters of the noise component are known, conventional identification and diagnostic methods may be used to estimate the signal component. In fact, these parameters are unknown and have to be estimated. Scott, Smith and Jones (1977) discuss two basic approaches. A direct designed based method follows the conventional sample survey approach by estimating lag error covariances directly from data on the sample units from which signal extraction weights can be derived. The second approach specifies a model of the aggregate series, e_t, with identifying restrictions that incorporate known features of the sample design.

These two approaches have their own particular strengths and weaknesses. As Bell and Hillmer (1987b) point out, the design based approach has the advantage that few constraints need to be imposed on the covariance structure of e_t. However, there are many practical problems involved with this approach, one of the most important being the availability of the micro data on individual sample units. The time series approach to modeling e_t has the advantage that it does not require these data. However, the error component cannot be identified from aggregate sample data without imposing restrictions.

Little research has been conducted using either approach for CPS data. Hausman and Watson (1985) developed an ARMA (1, 15) model of the error process for the national teenage unemployment rate series that incorporated the CPS 4–8–4 rotation design and the compositing procedure. An experimental application of the design based approach was performed by Bell and Hillmer (1987b) using the teenage data, but for a different time period. They developed an ARMA (1, 1) model as an approximation to the design based autocovariances estimated by Train, Cahoon and Makens (1978).

Producing covariance estimates for a survey as complex as the CPS is costly, requiring the availability and processing of a large amount of micro data. Currently, not all of the data we need are available, but will be forthcoming in the near future. The last section of this paper discusses our plans for developing CPS error models.

Because of these difficulties in developing designed based sample error covariances, we decided to initially fit the regression equations without attempting to estimate the individual effects of the errors in equation and the sample error. If the two component errors are ARMA processes, then their sum will also be an ARMA process (Granger and Morris, 1975). If

$u_t \sim \text{ARMA}(p_u, q_u)$, and $e_t \sim \text{ARMA}(p_e, q_e)$,

then,

$w_t = u_t + e_t \sim \text{ARMA}(p, q)$

where $p \le p_u + p_e$, $q \le \max(p_u + q_e, p_u + q_e)$. The KF may then be used to extract the regression component and the aggregate disturbance.

4.0 Unemployment Rate Models

4.1 Explanatory Variables

A common core of explanatory variables have been developed for the 40 State unemployment rate models. Each of the State models is based on two non-CPS data sources – – unemployment insurance claims developed from the Federal-State UI system and the Current Employment Statistics survey, a payroll survey of nonagricultural employment. These two data sources were the major inputs used in the Handbook methodology to prepare State estimates since the early 1960s.

To control for important cyclical and seasonal labor force movements not accounted for by the UI and CES data, variables have been constructed from selected CPS data in such a way as to reduce, if not eliminate, the influence of sample error. While this raises an "errors-in-variables" issue, it differs from the classical case in that our focus is on estimating the unobserved true value of the dependent variable rather than the coefficients. It may be profitable to use more State specific CPS data as explanatory variables, but to do so will require a model that explicitly accounts for errors in variables.

Monthly State data on the number of insured unemployed are the only source of current information on unemployment that are collected independent of the CPS. The nature of these data provide a starting point for the development of the rate model.

Insured unemployment data represent a complete count of the number of workers who are filing for UI benefits. Each State administers its own separate program subject to certain federal requirements (Blaustein, 1979). In general, benefits are paid only to workers who were laid off and meet certain State specific monetary and nonmonetary eligibility standards. In contrast, the concept of unemployment used in the CPS includes all persons who did not have a job during the survey week and who were looking for work or on layoff and waiting to be recalled to work. In terms of size, the most important groups of unemployed not included are shown in table 1. Of the three major ways persons become unemployed – – job loss, job leaving and labor force entry – – UI data essentially cover only a portion of one – – job losers.

Entrants account for the largest portion of unemployed not covered by UI. Because most have spent a substantial portion of their time prior to becoming unemployed outside the labor force, they will not have enough recent job experience and earnings to qualify for UI benefits. Job leavers who quit their jobs to search for other jobs generally are not eligible for benefits, at least for a period of time.

If the relative size of the unemployed not collecting UI benefits was stable over time, the claims rate would be an excellent proxy for the total unemployment rate. In fact, the labor market is characterized by large cyclical and seasonal shifts in the distribution of unemployment, particularly between job losers and entrants.

First, let us consider the seasonal movements in the distribution of unemployment. The most important phenomenon is the very different seasonal pattern of entrants and job losers. Youth and women make up the largest proportion of entrants. Youth unemployment shows a volatile seasonal pattern, reflecting the cycle of entry and exit from the labor force related to the school year. In contrast, job loss, the most common reason for adult male unemployment, reflects a seasonal layoff rehire pattern dominated by the annual production cycles of such industries as automobiles and construction.

Table 2 illustrates the difference between seasonal patterns (average of the 40 States) of the CPS entrant and job loser rate and their net effect on the total rate. Also shown is a typical seasonal pattern for the insured unemployed. The table entries are seasonal factors where a value greater (less) than 100 indicates a month of higher (lower) than average unemployment. The entrant rate is lower than average during the winter and higher than average in the summer, while the loser rate shows just the opposite pattern. The seasonal highs for each group have a strong influence on the total rate.

Of the three major categories of unemployed, job losers and entrants are numerically the most important, as, at different phases of the business cycle, they may account for as much as half of the total unemployed. Job leavers are less important quantitatively, usually accounting for about 15% of the total, but do have their own distinct seasonal pattern with highs in the late summer and early fall. Not surprisingly, the seasonal pattern of the claims rate closely follows job losers and does not reflect the influence of entrants or job leavers.

Next, we consider changes in the distribution of unemployment related to business cycles (see table 3). During recessions, as labor demand falls, layoffs rise. We would thus expect the claims rate to be a good cyclical indicator. While it is, there are several reasons why the claims count may not fully reflect the cyclical behavior of job losers. Towards the latter stages of a recession, the duration of spells of unemployment lengthen and the number of workers exhausting their UI benefits increases. Also, once reemployed, it may take some time for these workers to build up sufficient wage and employment credits to qualify for benefits during subsequent spells of unemployment. Most importantly, there appears to have been a secular decline in the UI coverage of unemployed workers. This is illustrated by table 3. Instead of rising during recession years, UI coverage declined. Quantitative research by Burtless and Vroman (1984) and Corson and Nicholson (1988) indicate that most of this decline was unrelated to economic or demographic changes and appears to have occurred because potential UI recipients never applied for benefits. Changes in public policy may have been the primary cause.

While the above discussion describes typical State behavior, there are important inter-State differences that must be accounted for in any modeling effort. In terms of UI data, there are variations in State eligibility requirements, benefit durations and administrative practices that have a major impact on UI coverage. This along with real differences in cyclical and seasonal behavior will cause regression coefficients and model parameters to differ across States. The diversity in UI coverage is illustrated by table 4. This table presents the means of annual averages of UI claims as a percent of total CPS unemployment by State, the coefficient of variation across years within a State and the minimum and maximum annual averages.

The general form of the regression component of the rate model is given below:

$$\begin{aligned}\text{Unemployment Rate} &= \text{Intercept}_t \\ &+ \beta_{1t}\ \text{Claims Rate} \\ &+ \beta_{2t}\ \text{Employment-to-Population Ratio (EP)} \\ &+ \beta_{3t}\ \text{Entrant Rate}\end{aligned} \tag{8.0}$$

where,

Claims Rate = (continued claims w/o earnings / CES employment) * 100

EP Ratio = (CPS employment / CPS 16+ population) * 100

Entrant Rate = (CPS entrant employed / (CPS entrants + CPS employment)) * 100.

As discussed above, the claims rate provides a measure of the relative size of those job losers who are collecting UI benefits. The EP ratio accounts primarily for those job losers not included in the claims counts. Given a relatively fixed labor force participation rate for experienced workers, their unemployment over the business cycle will be inversely related to the EP ratio. Although this variable is also sensitive to shifts in labor supply, its coefficient is negative, reflecting the dominance of labor demand fluctuations.

The EP ratio also can pick up seasonal fluctuations in labor demand that affect the number of unemployed job leavers. From a seasonal peak in the summer, the EP ratio declines in the fall. During this period, a large number of seasonal jobs come to an end, and unemployment due to job leaving rises.

In most States, the EP ratio was computed using CPS employment data in the numerator. While State CPS employment data are subject to sampling error, their coefficients of variation are five to six times smaller than the State CPS unemployment estimates. In a few States, CES data were used to compute the EP ratio. For most States the CES has a somewhat different seasonal pattern from the CPS employment measure, and is not likely to be as highly correlated with the "true" unemployment rate.

The effect of labor supply shifts on the unemployment rate, particularly on a seasonal basis, is primarily captured by the CPS entrant rate variable. To reduce the effect of sampling error, this variable is computed for a geographic area larger than the State (national, Census region or Census division level). In some cases, a 3-month moving average of the State CPS entrant rate is used. Seasonal shifts in the entrant rate coefficients are added in a few States, primarily to control for the May to June increase in entry of students into the labor force.

4.2 Estimation of State Models

For each of the 40 States, models were fitted to the CPS monthly unemployment rate series for the period 1976 to 1987. As previously discussed, the CPS sample data are represented in signal plus noise form as,

$$Y_t = \theta_t + e_t$$

with the signal represented by,

$$\theta_t = X_t\beta_t + u_t$$

where the coefficients follow a random walk,

$$\beta_t = \beta_{t-1} + v_{\beta t}.$$

Since we do not attempt to estimate the separate influences of u_t and e_t, the observed series is written as,

$$Y_t = X_t\beta_t + w_t$$

where $w_t = u_t + e_t \sim \text{ARMA}(p, q)$.

Let S_{wt} be the state vector form of w_t with order equal to max (p, q + 1). Then, the state space model has the following transition equations

$$S_t = \begin{bmatrix} \beta_t \\ S_{wt} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & T_w \end{bmatrix} \begin{bmatrix} \beta_{t-1} \\ S_{wt-1} \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & \Gamma_w \end{bmatrix} \begin{bmatrix} v_{\beta t} \\ v_{wt} \end{bmatrix}$$

with observation equation

$$Y_t = [X_t \; 1 \; 0 \cdots 0] \, S_t.$$

The parameters of this system are:

$\text{Cov}(v_{\beta t}) = \text{Diag}(\sigma_{\beta_1\beta_1}, \cdots, \sigma_{\beta_k\beta_k})$

$\text{Var}(v_{wt})$

T_w contains p AR parameters

Γ_w contains q MA parameters.

These parameters are estimated using the innovation form of the likelihood function (Schweppe, 1965). If the white noise disturbances $v_{\beta t}$ and v_{wt} are normally distributed, then the one–step ahead prediction errors or innovations, $\tilde{Y}_t$'s, are also independent N $(0, f_{t/t-1})$ random variables.

It follows that the joint probability of the observed sample data may be expressed as the product of the individual innovation densities. If the state vector contains ℓ nonstationary elements, then this joint density may be expressed as conditional on the first ℓ observations and the log of the likelihood as a function of the unknown parameters, denoted by, Ω, is within a constant,

$$L(\Omega) = -\frac{1}{2}\left[\sum_{t=\ell}^{n} \ln f_{t/t-1} + \frac{Y_t^2}{f_t}\right].$$

Given Ω, the KF recursions are used to evaluate $L(\Omega)$ using the first ℓ observations to compute starting values $S_{\ell+1/\ell}$ and $P_{\ell+1/\ell}$. The parameter space must then be searched to locate the maximum value of the likelihood. In general, this is a difficult nonlinear optimization problem. Initially, we simplified matters by specifying,

$$\text{Cov}(v_{\beta t}) = q\,D$$

where q is a scalar and D a diagonal matrix of prior specified constants, either all ones or ratios of the variances of the coefficients to the mean square errors as estimated from a fixed coefficient model. We also started with a first order AR model for w_t. This reduced the problem to the estimation of only two parameters. A coarse grid search was then performed to provide rough estimates of the degree of variation in the coefficients and the AR parameter value. These estimates, in some instances, were further refined by using them as starting values for the EM–scoring algorithm developed by Watson and Engle (1983). This algorithm allows for a general autoregressive structure for both coefficient change and the observation error.

Good initial values for the parameters are very important for obtaining convergence within a reasonable number of iterations. The grid search produced small values for the coefficient variances, $\sigma_{\beta_i\beta_i}$. The standard deviation of the coefficients was about .6 percent of the standard deviation of the observation error. Using these as starting values, the EM algorithm generally converged within 3 to 6 iterations. Using large initial values for these variances required over 40 iterations with convergence to a lower value of the maximum likelihood function. Diagnostic testing generally did not reveal inadequacies in the first order AR specification of the observation error.

There are a number of different ways to initialize the KF, as discussed by Harvey and Phillips (1979), Ansley and Kohn (1985), and Bell and Hillmer (1987a). While we used the information filter (which is not suitable for certain types of ARMA models) we found no problems with using the regular KF equations. Given calculations done in double precision, the KF produced results identical to the information filter. We tested the KF further by setting the coefficient variances to zero and compared the results with conventional regression software and again found no significant differences.

4.3 Diagnostic Tests

Once the models are estimated, the innovation series – – the difference between the current sample estimate and the best prediction of its value produced by the KF – – provides a natural check on overall goodness of fit. If the parameters of the state space model are known, then the standardized innovations will be normal and independently distributed with unit variance,

$$\tilde{y}_t = (Y_t - Y_{t/t-1}) \,/\, \sqrt{H_t\, P_{t/t-1}\, H_t'} \sim \text{NID}(0, 1).$$

As discussed below, we test for a wide variety of departures from these properties. Of course, the test results should be treated as exploratory in nature rather than formal tests of significance since the state space parameters are estimated and repeated testing is performed on the same data. Nevertheless, it is reasonable to expect that strong departures from model assumptions would imply misspecification.

(1) Autocorrelation

A variety of tests were performed based on both correlogram and frequency domain approaches. The correlogram was computed as

$$\hat{r}_\ell = \sum_{t=s+1+\ell}^{n} \tilde{y}_t \tilde{y}_{t-\ell} / \sum_{t=s+1}^{n} \tilde{y}_t^2$$

where,

s = number of state variables
n = total number of observations.

In inspecting the correlograms for each model we used $\pm 2 / \sqrt{n-s}$ as a rough confidence interval around $\hat{r}_\ell$. The Durbin–Watson and the Ljung–Box version of the portmanteau test were used to supplement the correlogram analysis.

$$DW = \sum_{t=s+2}^{n} (\tilde{y}_t - \tilde{y}_{t-1})^2 / \sum_{t=s+1}^{m} \tilde{y}_t^2$$

$$LB(m) = (n-s)(n-s+2) \sum_{\ell=1}^{m} \hat{r}_\ell^2 / (n-s-\ell) \sim \chi^2_{m-p-q}$$

where p and q represent the order of the ARMA model fitted to the observation errors.

Frequency domain methods are useful for revealing certain types of non-independence in the computed innovations that are less clearly detectable by the correlogram. In particular, cyclical properties may not be well characterized by the correlogram. Since the labor force is a highly seasonal series, we are particularly interested in detecting correlations in the innovations that are related to seasonal errors. We tested specifically for departures from white noise behavior at the seasonal frequencies. The statistic for testing this hypothesis is the ratio of the spectral density of the innovation to that for a white noise process.

The spectral density for the innovations was computed as a weighted average of its periodogram ordinates, using a symmetric three-point weighting scheme. The spectral density for the white noise process is proportional to the total variance of the innovation. The test statistic is computed as

$$v \frac{S(f_i)}{S_0(f_i)} \simeq \chi^2_v \quad \text{for } f_i = 1/12, 1/6, 1/4, 1/3, 1/2.4$$

where,

$$S(f_i) = \sum_{j=-1}^{1} w_j P(f_i),\ w_j = \frac{2-|j|}{4},\ v = 2\left[\sum_{j=-1}^{1} w_j^2\right]^{-1},$$

$$S_0(f_i) = \frac{\chi^2_{\tilde{y}}}{2\pi}.$$

(2) Heteroscedasticity with time

Because there have been numerous changes in sample design and large increases in labor force size over the sample period, change in the variance of the standardized innovations is a real possibility. This is tested with the following statistic,

$$H = \sum_{t=n-s-m+1}^{n-s} \tilde{y}_t^2 \ / \sum_{t=s+1}^{s+m} y_t^2 \sim F_{m,m},\quad m = \frac{n-s}{3}$$

where the middle third of the observations are dropped to more easily detect differences between the average variances of the first and later part of the series. We also examined plots of the standardized innovations over time for evidence of heteroscedasticity.

(3) Normality

Potential departures of the innovations from normality are tested using the Bera–Jarque (1987) statistics based on measures of skewness (b_1) and kurtosis (b_2),

$$\lambda = T\left[\frac{(b_1)^2}{6} + \frac{(b_2-3)^2}{24}\right] \simeq \chi^2_2,\quad T = \text{total number of observations.}$$

This is a joint test of whether or not the estimates of skewness b_1 and excess kurtosis, $(b_2 - 3)$, are significantly different from zero.

(4) Post–Sample Tests

The innovation series produced by the KF are particularly well–suited to testing the adequacy of the model to make accurate predictions outside the set of data used in its construction (Harvey, 1981). We divide our data into two parts, a sample period (1976–86) and a post–sample period (1987). A test that prediction errors are greater in the post–sample period is given by

$$\sum_{j=1}^{\ell} \left[\tilde{y}^2_{T+j} / \ell\right] / \sum_{t=s+1}^{T} \left[\tilde{y}_t^2 / (T-s)\right] \sim F_{\ell,T-s}$$

where t > T is the post–sample period.

In addition, it is useful to test for prediction error bias. A standard t statistic is used for this purpose,

$$t_{t-s-1} = \sqrt{n-s}\ \bar{\tilde{y}} / s_{\tilde{y}}$$

where,

$$\bar{\tilde{y}} = \sum_{t=s+1}^{n} \tilde{y}_t / (n-s)$$

$$s^2_{\tilde{y}} = \sum_{t=s+1}^{n} (\tilde{y}_t - \bar{\tilde{y}})^2 / (n-s-1).$$

4.4 Model Performance

About 60 to 70 percent of the total variance in the monthly CPS series is attributable to the regression mean with the remainder due to the aggregate noise term. We have not yet developed mean square error estimates for the signal. To do so will require information on the CPS sample error covariances.

The time varying regression mean looks considerably smoother than the CPS series. This can be seen by visually examining the plots of the various series. However, the highly seasonal nature of the labor force makes this visual comparison somewhat misleading. A useful rough summary measure of smoothness of a series is to decompose it into trend–cycle, seasonal and irregular components and compute the percent of the variance of monthly change due to the irregular. This smoothness measure was computed for both the model and CPS series. For the unemployment models, the median percent irregular (30 percent) is about half that of the corresponding CPS series.

Based on the diagnostic tests, the 40 models appear to fit the systematic underlying movements in the CPS fairly well. There are, however, a number of areas that need improvement. Table 5 shows the number of models that did not pass one or more of the tests. The major problems identified with the unemployment models were high order autocorrelation in 11, and heteroscedasticity in 9 of the 40 models. The primary source of the first problem appears to be the way in which some of the CPS explanatory variables were computed. Of the 11 models with positive Ljung–Box tests, 10 included CPS explanatory variables (either EP or entrant rate) that had been computed as three month moving averages of the original data. If the smoothing of these variables is eliminated, then 8 of these States would pass the Ljung–Box tests. The reason for using the moving averages was to dampen large irregular movements in the series. A more appropriate approach is to directly model the CPS explanatory variables as functions of their past values and directly incorporate them into the KF model. This approach is discussed in the last section of this paper.

The heteroscedasticity identified in 9 of the unemployment rate models may in part be a reflection of changing variances in the CPS sample error. In general, the whole modeling approach could benefit by explicitly accounting for specific CPS sample design features, including heteroscedasticity of CPS sample errors.

5.0 Current and Future Research

There is an ongoing program of research to develop further improvements to the time series models. This section discusses two major areas, modeling the CPS sample error and the errors–in–variables issue that arises in using CPS data as explanatory variables.

5.1 Modeling CPS Sample Error

As discussed in section 2.0 of this paper, information on the structure of the CPS sample error is necessary in order to decompose the composite regression disturbance term into its sample error and model error components. Given CPS error variances and lag covariances, ARMA models can be developed to approximate the time series behavior of the sample error. Treating the ARMA coefficients as known parameters of the state space system, standard time series diagnostic tools may be used to model the errors in equation disturbances.

Currently, the major constraint to the development of sample error models is the lack of appropriate sample data. While the Census Bureau has routinely produced State variance estimates, based on generalized variance functions, since the mid–1970s, lag error correlations are more costly to obtain. As part of the rewrite of the CPS production system, BLS will begin receiving replicate weight files that can be used to compute correlations at varying levels of geographic detail.

This will be an extensive project which requires processing of large amounts of CPS micro data. In the meantime, we are planning to use the CPS generalized variance estimates currently available to model the heteroscedastic structure of the error terms and experiment with alternative covariance structures for the sample error.

5.2 Errors in Variables

The CPS data used as explanatory variables in the unemployment rate model are themselves the sums of their true stochastic values plus measurement error. Thus, we have signal plus noise terms for both the dependent and regressor variables,

$$Y_t = \theta_{Yt} + e_{Yt}$$
$$\theta_{Yt} = X_t\beta_t + u_{Yt}$$
$$X_t = \theta_{Xt} + e_{Xt}\,.$$

For the fixed coefficient case, a well–known result is that random measurement errors induce an asymptotic bias in the coefficient estimators that is a function of the signal to noise variance ratio in the observed X series. The importance of this bias depends upon the purpose of the regression analysis. If it is used for predicting the dependent variable, on the basis of the observed X, then the errors–in–variables problem is less severe. In the analogous fixed coefficient case, it can be shown that the prediction is unbiased, provided the noise is stationary and retains those statistical characteristics it possessed in the data used to obtain the estimate (Johnston, 1963, Fuller, 1987).

The approach used in this paper is to condition on the X variables. If Cov (e_{Xt}, e_{Yt}) were zero we could consistently estimate β, assuming it is a fixed parameter, as the empirical relationship between θ_{Yt} and X_t. However, given that Y_t and some X_t come from the same survey, this covariance will not be zero, although in practice it may be small. For example, in the State unemployment rate model the CPS entrant rate is generally computed for a Census division, regional or national level. The CPS EP variable, while State specific, has a relatively small sample error.

There are several approaches we could take to the errors–in–variables problem. We could drop the CPS explanatory variables and retain only the non–CPS variables and add stochastic trend and seasonal variables to control for behavior in the CPS not well reflected in the non–CPS data. Another approach is to retain the CPS explanatory variables but model their stochastic structure. The θ_{Xt}'s are clearly not independently distributed over time. This existence of an autocorrelation structure implies a certain degree of smoothness that could be exploited to solve the errors–in–variables problem. One way to do this is to model the observed X series in state space form,

$$X_t = \theta_{Xt} + e_{Xt} = H_X\, S_{Xt}$$

$$S_{Xt} = T_X\, S_{X,t-1} + \Gamma\, v_{Xt}$$

$$S_{X,t/t} = T_X\, S_{X,t/t-1} + K_{Xt}\, \tilde{X}_t$$

$$K_{xt} = \text{Cov}\,(S_{X,t/t-1}, X_{t/t-1})\,(\text{Var}\,(X_{t/t-1}))^{-1}$$

$$\tilde{X}_t = X_t - H_X\, S_{t/t-1}$$

where,

S_{Xt} = a state vector

T_X = transition matrix

v_{Xt} = vector of white noise disturbances

$S_{X,t/t-1}$ = conditional expectation of S_{Xt} given $X_{t-1}, \cdots, X_1$

K_{Xt} = gain of the Kalman filter

$\tilde{X}_t$ = one–step ahead prediction error or innovation.

It follows from the above that the observed series may be written as

$$X_t = H_X\, S_{X,t/t-1} + \tilde{X}_t$$
$$= X_{t/t-1} + \tilde{X}_t.$$

Since $X_{t/t-1}$ is a function of past X's only it is uncorrelated with X_t and also contemporaneously uncorrelated with measurement error in Y_t. Thus $X_{t/t-1}$ may be used as an instrumental variable to replace X_t in the regression equation,

$$Y_t = X_{t/t-1}\, \beta_t + w_t.$$

This basic approach of using the autoregressive structure to smooth the X variables can be extended in a number of directions as suggested by Mehra (1976) and Eltinge (1987). The observation and transition equations for the CPS X–variables could be added to the state space equations for the dependent variable so that the autocorrelation of the X's could be estimated simultaneously with all the other parameters using maximum likelihood. If, in addition, we want to account for the sample error structure in the CPS X–variables we could construct a multivariate signal–extraction model. Time varying coefficients could be retained, but then the measurement equations would be nonlinear. The basic KF structure could be preserved by using an extended KF.

Acknowledgements

The author thanks Art Dempster, Steve Miller, Jack Bregger, Art Hughes and Tony Barkume for their helpful comments and Thomas Evans for assistance in the preparation of the text and tables.

The views expressed in this paper are those of the author and do not necessarily represent the policies of the Bureau of Labor Statistics.

REFERENCES

Ansley, C.F., and Kohn, R. (1985), "Estimation, Filtering, and Smoothing in State Space Models with Incompletely Specified Initial Conditions," *Annals of Statistics*, 13, 1286–1316.

Bell, W.R., and Hillmer, S.C. (1987a), "Initializing the Kalman Filter in the Nonstationary Case," Bureau of the Census Statistical Research Division Report Series, #CENSUS/SRD/RR–87/33.

——— (1987b), "Time Series Methods for Survey Estimation," Bureau of the Census Statistical Research Division Report Series, #CENSUS/SRD/RR–87/20.

——— (1989), "Modeling Time Series Subject to Sampling Error," Bureau of the Census Research Division Report Series, #CENSUS/SRD/RR–89/01.

Blaustein, S.J. (1979), paper prepared for the National Commission on Employment and Unemployment Statistics.

Bureau of Labor Statistics (1988), *BLS Handbook of Methods*, Bulletin 2285, Washington, D.C.: Author.

Bureau of the Census (1978), *The Current Population Survey: Design and Methodology*, Technical Paper 40, Washington, D.C.: Author.

Burtless, G., and Vroman, W. (1984), "The Performance of Unemployment Insurance Since 1979," paper presented at the December 1979 meeting of the Industrial Relations Research Association, Dallas, Texas.

Corson, W., and Nicholson, W. (1988), "An Examination of Declining UI Claims During the 1980s," Mathematica Policy Research Draft Report.

Eltinge, J.L. (1987), "Measurement Error Models for Time Series," in *Proceedings of the Business and Economic Section, American Statistical Association*, pp. 88–94.

Engle, R.F., and Watson, M.W. (1974), "Application of Kalman Filtering in Econometrics," Harvard Institute of Economic Research Discussion Paper #1187.

Ericksen, E.P. (1974), "A Regression for Estimating Population Changes of Local Areas," *Journal of the American Statistical Association*, 69, 867–875.

Fuller, W.A. (1987), *Measurement Error Models*, New York: John Wiley.

Gelb, A. (ed.) (1974), *Applied Optimal Estimation*, Cambridge, MA: MIT Press.

Granger, C.W.J., and Morris, M.J. (1976), "Time Series Modeling and Interpretation," *Journal of the Royal Statistical Society*, Ser. A, 139, 246–257.

Harvey, A.C. (1981), *Time Series Models*, Oxford: Philip Allan.

Harvey, A.C., and Phillips, G.D.A. (1979), "Maximum Likelihood Estimation of Regression Models with Autoregressive–Moving Average Disturbance," *Biometrika*, 66, 49–58.

Hausman, J.A., and Watson, M.W. (1985), "Errors in Variables and Seasonal Adjustment Procedures," *Journal of the American Statistical Association*, 80,531–540.

Jarque, C.M., and Bera, A.K. (1987), "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, 55, 163–172.

Johnston, J. (1963), *Econometric Methods*, New York: McGraw–Hill.

Los, C.A. (1985), "Regression Analysis and Diagnosis by an Evolutionary Parameter Estimator," Federal Reserve Bank of New York Research Paper.

Maybeck, P.S. (1979), *Stochastic Models, Estimation, and Control* (Vol. 2), Orlando: Academic Press.

Mehra, R.K. (1976), "Identification and Estimation of the Error–in–Variables Model in Structural Form," Mathematical Programming Study, 191–210.

Schweppe, F. (1965), "Evaluation of Likelihood Functions for Gaussian Signals," *IEEE Transactions on Information Theory*, 11, 61–70.

Scott, A.J., and Smith, T.M.F. (1974), "Analysis of Repeated Surveys Using Time Series Methods," *Journal of the American Statistical Association*, 69, 674–678.

Scott, A.J., Smith, T.M.F., and Jones, R.G. (1977), "The Application of Time Series Methods to the Analysis of Repeated Surveys," *International Statistical Review*, 45, 13–28.

Train, G., Cahoon, L., and Makens, P. (1978), "The Current Population Survey Variances, Inter–Relationships, and Design Effects," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 443–448.

Table 1. Major Groups of Unemployed Not Counted in Unemployment Insurance Data

1. Job Losers in following categories:	
a. Exhaustees –	workers who have exhausted their benefit entitlements.
b. Monetary Ineligibles –	workers with insufficient prior employment or earnings to meet State eligibility requirements.
c. Delayed and Never Filer –	eligible workers who do not file for benefits at the start of their unemployment spell.
2. Labor Force Entrants –	workers who prior to their current spell of unemployment were outside the labor force.
3. Job Leavers –	workers who left their previous job but continue to look for another job.

Table 2. Average of Unemployment Seasonal Factors for 40 State Aggregate, 1979–85

	CPS Unemployment Rates				
Month	Total	Entrants	Job Losers	Job Leavers	UI Claims Rate
Jan	110.6	99.1	120.9	104.3	131.2
Feb	110.9	98.1	126.1	100.7	133.8
Mar	105.3	96.2	115.7	95.5	120.8
Apr	97.0	90.7	103.5	91.3	104.0
May	93.2	96.4	91.0	92.8	91.0
Jun	106.0	127.2	89.6	93.6	85.8
Jul	98.7	106.0	90.8	101.0	95.9
Aug	98.5	102.3	92.1	110.7	90.2
Sep	95.5	102.7	84.7	112.8	77.6
Oct	93.6	98.3	88.0	107.0	79.3
Nov	94.9	94.5	93.5	101.3	87.7
Dec	95.9	88.4	104.1	89.5	102.4

NOTE: Seasonal factors were computed from X-11. The denominator of a CPS rate is the sum of CPS employment plus unemployment for the specific category. The denominator for the UI claims rate is total CES employment.

Table 3. Relative Size of Categories of Unemployment, Average of 40 States by Year

		Percent of Total CPS Unemployment				
Year	U.S. Unemp. Rate	UI Claims as % of CPS Job Losers	Claims	Job Losers	Entrants	Job Leavers
76	7.7	93.1	36.9	43.1	41.9	14.8
77	7.1	86.5	34.5	41.3	43.6	14.9
78	6.1	87.6	31.4	37.7	46.1	16.2
79	5.8	87.5	32.8	39.4	44.4	16.1
80	7.1	76.7	35.8	47.6	38.3	14.0
81	7.6	68.1	31.9	48.8	38.4	12.7
82	9.7	61.8	34.3	56.3	34.5	9.1
83	9.6	50.7	27.7	55.1	35.5	9.2
84	7.5	52.2	25.7	50.1	38.9	10.9
85	7.2	56.6	26.9	48.3	40.1	11.5
86	7.0	60.3	27.9	47.7	38.9	13.3
87	6.2	57.0	25.3	45.6	40.2	14.0
88	5.5	62.3	26.4	43.9	40.4	15.5

Table 4. UI Claims as a Percent of Total CPS Unemployment by State, 1976–87

State	UI as a Percent of CPS	Coefficient of Variation over Years	Minimum Value	Maximum Value
AL	25.5	23.7	17.5	37.5
AK	57.1	21.2	43.1	83.0
AZ	24.3	13.8	19.8	29.5
AR	28.1	21.6	20.2	38.5
CO	23.6	10.9	20.2	28.1
CT	37.5	17.0	29.3	48.9
DE	29.9	17.1	22.8	38.0
DC	26.6	10.7	21.9	30.3
GA	25.8	15.0	21.2	33.4
HI	32.9	9.5	29.0	39.8
ID	31.6	15.4	24.7	39.5
IN	24.4	17.5	19.4	33.5
IA	30.3	20.9	22.6	42.1
KS	35.5	13.6	29.5	45.9
KY	29.2	29.7	17.2	40.0
LA	27.9	20.6	20.0	37.6
ME	35.8	11.2	28.7	41.1
MD	28.5	12.2	23.6	34.0
MN	33.7	18.7	24.0	44.4
MS	26.1	19.7	20.1	35.3
MO	32.8	19.5	24.3	44.2
MT	32.2	21.6	23.1	44.2
NE	30.0	18.7	20.7	43.9
NV	34.7	19.7	25.3	48.2
NH	26.4	17.4	19.2	34.7
NM	24.2	10.5	20.4	28.1
ND	31.1	14.7	24.3	36.7
OK	27.5	21.7	19.4	39.0
OR	37.6	16.7	28.8	48.4
RI	51.2	12.7	39.5	61.4
SC	26.3	18.1	19.8	33.9
SD	22.5	30.6	14.7	35.0
TN	27.2	26.2	18.1	41.0
UT	31.4	25.7	19.9	41.1
VT	40.0	9.7	33.9	46.3
VA	17.6	20.1	13.3	24.7
WA	36.1	15.7	30.1	49.1
WV	32.8	24.7	21.1	42.7
WI	34.2	24.1	24.8	47.1
WY	28.7	28.7	20.7	49.7
All	31.0	23.2	17.6	57.1

Table 5. Summary of Diagnostic Checks: States with Significant Test Values at .05 Level

Test	Number of States
Durbin Watson	1
Ljung–Box [–12]	7
Ljung–Box [–24]	10
Spectral Density	
12 month frequency	9
6 Month frequency	3
4 month frequency	3
3 month frequency	5
Heteroscedasticity	9
Bera–Jarque Normality	4
Post–Sample Prediction	1
Post–Sample Bias	0