Predicting the National Unemployment Rate that the "Old" CPS Would Have Produced

Richard Tiller and Michael Welch, Bureau of Labor Statistics
Richard Tiller, Bureau of Labor Statistics, Room 4985, 2 Mass. Ave., N.E., Washington, DC 20212-0001

KEY WORDS: Time series forecasting, Structural Models

Abstract: In January 1994, the introduction of the redesigned Current Population Survey (CPS) questionnaire and automation of collection procedures was expected to affect most labor force estimates. To help evaluate the change in the unemployment rate attributed to these revisions, time series models were used to extrapolate the pre- 1994 series to predict the unemployment rate estimates for 1994.

I. Introduction

Beginning with January 1994 data, the Current Population Survey (CPS) introduced new data collection procedures and population controls based on the 1990 census, adjusted for census undercount. These new procedures may result in substantial changes in many labor force series, including the national unemployment rate. In order to address the issue of comparability between the "old" and "new" series for various groups of data users, time series models were developed by the Bureau of Labor Statistics to predict what the national unemployment rate would be during the early months of 1994 under the "old" CPS data collection procedures and population controls based on the 1980 census.

The model uses the historical relationships between CPS data and unemployment insurance claims for the CPS reference week and employment from the Current Employment Statistics Survey (the BLS payroll survey of business establishments). The model was fitted to data from January 1976 through December 1993, the last month for which official estimates were made using the "old" data collection procedures. As soon as data are validated from the new parallel survey, which will use the "old" CPS methods, these data will be incorporated into a model to estimate what the monthly unemployment rate would have been had the "old" survey been continued. The new model and sample-based estimates then may replace the projections described in this paper.

This report discusses background of the CPS; gives a brief description of the data used in the models; presents the model and examines test statistics relevant to assessing its performance; predicts the unemployment rate, not seasonally adjusted, that would have been produced had the "old" survey been continued in 1994; describes the methods used to seasonally adjust the model-based prediction; and offers caveats concerning the predictions. Additional technical detail is provided in the complete paper.

II. Background

The CPS is a monthly probability sample survey of about 60,000 households, conducted by the Bureau of the Census for the Bureau of Labor Statistics (BLS). Beginning with the January 1994 interview, the CPS is conducted using a new questionnaire in a completely computer-assisted environment. The Bureau of the Census and the BLS tested the new procedures for 18 months (July 1992 - December 1993) on a separate, national-based probability sample of 12,000 households. The results of this parallel survey indicate that the CPS annual average unemployment rate would have been 0.5 percentage point higher in 1993 if the new approach had been used. Additionally, the introduction of 1990-based population controls raises the unemployment rate 0.1 percentage point more than that obtained from 1980-based population controls. Additional effects due to design differences are discussed in Kostanich and Cahoon[1].

To better understand the differences between the "old" and "new" methodology, we are switching the old CPS procedures to the parallel survey sample of 12,000 households (here in after "new parallel survey"). In other words, in January 1994, the CPS sample of 60,000 households began using the "new" methods, and the parallel survey sample of 12,000 households began using the "old" methods. Due to operational constraints, it was not possible to avoid this switch-over with its possible attendant effects on respondents and interviewers.

Although data are being collected using both the old and new collection methods, the official labor force estimates are based on the CPS using the new methods. We cannot provide the public with an immediate source of comparison between the "new" and "old" labor force estimates because the reliability of data from the new parallel survey may be low during the initial months, due to nonsampling errors associated with the start-up period that are beyond our control. As an interim measure, we developed a structural time series model to predict what the monthly national unemployment rate would have been had the "old" CPS been continued. This paper outlines the research conducted jointly by the BLS, the Bureau of the Census, and consultants from Iowa State University to develop this prediction.

III. Description of data

The data used for modeling the unemployment rate cover the period January 1976 through December 1993. These data consist of estimates of the civilian noninstitutional population and the unemployment rate from the CPS, estimates of employment from the Current Employment Statistics (CES) survey, and unemployment insurance continued claims counts provided by the Employment and Training Administration. The CPS and CES data are official BLS estimates obtained from the Bureau's LABSTAT database. Data are not seasonally adjusted, and levels are rounded to the nearest thousand.

The CPS data are composited and based on 1980 population controls. The CES data are final benchmarked up to March 1992, first benchmarked for the period April 1992 through April 1993, third closing for the period May 1993 through November 1993, and second closing for December 1993. Although the most recent CES data are subject to further revision, for the sake of consistency, we will not use data reflecting future revisions to reestimate our model. The unemployment insurance claims counts are the total number of regular state unemployment insurance claims filed during the week that includes the CPS reference week. These do not include claims paid under the Emergency Unemployment Compensation Act or earlier extended benefits provisions.

IV. The prediction model

A number of different time series models were fit to CPS unemployment rate data for the period January 1976 through December 1993 for a total of 216 observations. The alternatives considered were structural time series models with explanatory variables[2], multiple regression with autocorrelated disturbances[3], and univariate ARIMA models[4]. (See the appendix for more details.) A structural time series model was selected as the preferred model because of its goodness of fit to the historical data, forecasting performance, and ease of explanation.

The structural model is essentially a multiple regression that includes a trend and seasonal component and two explanatory variables as regressors. This model differs from the usual regression model in that the trend and seasonal components do not have a fixed functional form over the entire sample period but rather are allowed to vary smoothly over time. The model is given by

$$Y_t = \mu_t + \beta_1 CLR_t + \beta_2 CESEP_t + S_t + \varepsilon_t \ ,$$

where

Y_t = CPS unemployment rate for month t,

μ_t = time varying trend term,

$CLR_t = 100(UI_t/CESEM_t)$,

$CESEP_t = 100(CESEM_t/POP_t)$,

UI_t = unemployment insurance claims,

$CESEM_t$ = employment level from the CES,

POP_t = civilian noninstitutional population,

$\beta_{1,}\ \beta_2 =$ fixed regression coefficients,

S_t = the seasonal component, and

ε_t = a random disturbance (noise) term.

The two explanatory variables used in the model are the ratio of worker claims for unemployment insurance benefits to CES employment (CLR) and the ratio of CES employment to the estimated civilian noninstitutional population (CESEP).

The CLR and CESEP variables are included in the model because they are strongly correlated with the CPS unemployment rate, and are readily available on a timely basis. However, the variables do not explain a significant amount of variation in the CPS rate. A complete explanation would require a complex model with many variables. As an alternative to such a complex model, we add stochastic trend and seasonal components to capture both long-run movements and seasonal variation in the CPS unemployment rate that are not accounted for by the two regressors (CLR and CESEP). Note that in this model the seasonal component reflects the seasonal pattern in the unemployment rate not accounted for by the explanatory variables and thus it is not suitable for seasonally adjusting the unemployment rate.

The trend component, η_t, or time varying intercept, is represented as a nonstationary autoregressive process (random walk). That is, its current value is equal to its previous period value plus a random disturbance. Thus, the trend will change very smoothly over time, shifting up or down, with no persistent directional change. The magnitude of the change is determined by the variance of the disturbance term. Similarly, the seasonal component is specified as a nonstationary process consisting of the sum of six trigonometric terms with seasonal periodicities. Each of these components contains a random disturbance with a common variance. This allows the amplitude and phase of the seasonal pattern to change slowly over time, where the degree of change depends upon the size of the disturbance variance.

The effect of specifying the trend and seasonal components in the fashion just described is to discount past observations in the computation of these components. Thus, data from the 1990's are assumed to be more relevant for predicting the trend and seasonal components in 1994 than are data from the 1970's. The degree of discounting depends upon the size of the variances of the trend and seasonal components. These variances are determined empirically.

Table 1 presents the values of the estimated coefficients and t-ratios for the two explanatory variables, and monthly estimates of the trend and seasonal components for 1993. The trend has a large positive value, but is offset by multiplying the CESEP variable by its negative coefficient.

In the initial model estimation, the seasonal pattern was estimated to vary smoothly over time. A closer examination, as suggested by Wayne Fuller of Iowa State University, revealed that most of the change in the seasonal component was occurring in May and June, months when teenagers have a strong influence on labor force movements. There has been a secular decline in the relative size of this teenager group, which might explain the observed changes in the seasonal pattern. To test this possibility, a seasonal change variable for May and June expressed as a function of the percent of 16 to 19-year-olds to total population was introduced. When this variable was added to the model, the variance in the residual seasonal component was reduced to zero. While this had little effect on the final predictions, it did reduce the standard deviation of the prediction error by 15 percent.

The lower part of Table 1 presents evaluative statistics. The standardized one-step ahead prediction errors generated from the model were tested for autocorrelation, non-normality, and increasing variance over the 1993 sample period. The Q statistic is the portmanteau test for autocorrelations in the prediction errors up to 24 lags. This statistic has an asymptotic chi-squared distribution with 24 degrees of freedom. A value of about 40 or more would indicate significant autocorrelations. The normality test can be compared to a chi-square distribution with two degrees of freedom. A value higher than about six would indicate lack of normality. The variance test checks for larger prediction errors in the last third of the sample relative to the first third. This test statistic has an F distribution. The root mean square error (RMSE) is the standard deviation of the one-step-ahead prediction errors computed over the last year of the sample period. This statistic measures how well model predictions compare to actual observations. None of the diagnostics in table 1 suggests that the model is inappropriate.

Three alternative coefficients of determination (R^2, R_D^2, and R_s^2) are shown as measures of goodness of fit. The conventional R^2 is 1 minus the sum of squared prediction errors to the sum of squared deviations of the unemployment rate observations about the mean. It shows how much of the variation in the series is explained by the full set of model variables, including the time-varying intercept and the seasonal factors. The R_D^2 measure indicates how much of the variation in the first difference of the series can be explained by the model. The R_S^2 measure is even more stringent; it represents the share of the residual variation explained by the model after taking first differences and then subtracting seasonal means. This measure is considerably lower than the value for R^2. Nevertheless, the model makes a relatively large contribution to explaining the variation in the unemployment rate that remains even after trend and seasonal movements have been factored out of the series.

Table 1. Model Estimates and Evaluative Statistics

	Coefficients/components (T-ratios in absolute value)	
	CESEP[1]	CLR[2]
	-0.47 (6.9)	0.56 (7.3)
	Trend (1993)	Seasonal (1993)
Jan	32.58 (8.1)	-0.09 (1.5)
Feb	32.61 (8.1)	-0.09 (1.6)
Mar	32.53 (8.1)	-0.20 (4.6)
Apr	32.48 (8.0)	-0.35 (10.9)
May	32.40 (8.0)	0.08 (1.3)
Jun	32.40 (8.0)	0.46 (14.7)
Jul	32.31 (8.0)	0.07 (2.1)
Aug	32.20 (8.0)	-0.10 (2.5)
Sep	32.12 (7.9)	0.13 (2.8)
Oct	32.09 (7.9)	0.08 (1.9)
Nov	31.98 (7.9)	0.14 (3.4)
Dec	31.94 (7.9)	-0.12 (2.9)
	Evaluative statistics	
Q	12.83	
Normality	1.04	
Variance Test	1.20	
Rmse	0.17	
R^2	0.98	
R_D^2	0.85	
R_s^2	0.31	

Predictions

Table 2 presents the official unemployment rate estimates for 1993 with associated standard errors and 90 percent confidence intervals together with the predicted values for January through October 1994, their standard errors, and 90-percent confidence prediction intervals. The standard errors are computed from the model. The prediction intervals will become longer as the prediction period is extended.

The predicted rate is seasonally adjusted by using the implicit seasonal factors derived from the official rate estimates (discussed in detail later in this report). Approximate confidence intervals for the seasonally adjusted estimates are computed using the standard errors for the unadjusted data.

V. Seasonal adjustment procedure

The seasonally adjusted national unemployment rate from the CPS is produced by aggregating 12 independently adjusted series. The component series are: agricultural employment, nonagricultural employment, and unemployment, each for four sex-age groups (men 20 years and older; women 20 years and older; men 16 to 19 years; and women 16 to 19 years). Eight of these series are seasonally adjusted using multiplicative adjustment factors; the remaining four -- nonagricultural men and women aged 16 to 19 years, and unemployed men and women aged 16 to 19 years use additive adjustment factors.

The seasonal adjustment factors are generated using X-11 ARIMA software, and the factors for 1994 are given in the January 1994 issue of *Employment and Earnings*. Each of the 12 series is separately adjusted for seasonal variation. The series then are added to derive seasonally adjusted aggregate figures. The seasonally adjusted unemployment estimate is a sum of four seasonally adjusted unemployment components. The seasonally adjusted figure for the civilian labor force is a sum of eight seasonally adjusted civilian employment components and four seasonally adjusted unemployment components. The overall unemployment rate is derived by dividing the estimate of unemployment by the estimate of the civilian labor force.

The modeling described here yields an estimate of the unemployment rate, not seasonally adjusted. A seasonally adjusted rate was calculated by multiplying the unadjusted rate estimate by the ratio of the official January 1994 adjusted rate to the official January 1994 unadjusted rate. This approach seemed reasonable because analysis indicated that monthly differences between CPS and initial parallel survey unemployment rates were not affected by seasonal adjustment.

The official CPS unemployment rate, seasonally adjusted, for January 1994 is 6.7 percent, and the not seasonally adjusted unemployment rate is 7.3 percent. The ratio of the seasonally adjusted rate to the not seasonally adjusted one is, therefore, 0.9178. To obtain the seasonally adjusted prediction of the January 1994 unemployment rate that would have been produced by the "old" CPS methods, we multiply the not seasonally adjusted prediction of 6.9 percent by 0.9178. This gives us a seasonally adjusted prediction of 6.3 percent for January 1994.

VI. Caveats

It is important to note that the predicted estimates are based on historical relationships that may or may not carry over into the future. Specifically, it should be noted that no concurrent CPS data are used in the model to reflect the old CPS questionnaire and data collection methodology. This means that disturbances to the economy in early 1994 will not be reflected in the predictions, except as captured by the explanatory variables. In view of this, the predictions should be interpreted with caution, especially when the period is extended beyond January. As soon as data from the new parallel survey that replicates the "old" CPS methods have been validated, they will be incorporated into a model to estimate what the monthly unemployment rate would have been had the "old" survey been continued. These model and sample based estimates will then replace the projections described in the present report. Production of these estimates will continue, as we seek to help users better understand the relationship between the new, official series and the data derived from the "old" CPS.

APPENDIX: Description of the Modeling Methods

Three different approaches to time series models were used to estimate alternative forecasts of the CPS unemployment rate in 1994. These methods are based on the structural modeling approach[1]; autoregressive-integrated-moving-average (ARIMA) models[2]; and multiple regression models. The structural model provided the best alternative to satisfying the objective of multi-period forecasting with explanatory variables. This appendix provides further technical detail on the structural modeling method and then briefly addresses the regression model and ARIMA approaches considered.

Structural modeling with explanatory variables. This approach, as exemplified in the work of A.C. Harvey[3], explicitly models components known to exist in a time series, such as trends, seasonals, and irregulars. In univariate form, these models are closely related to ARIMA models, but do not include as wide a class of models as the G.E.P. Box-G.M. Jenkins approach[4]. When explanatory variables are added to this model, it is similar to a regression model. The general form of the model used in our application is described below.

Let

$$Y_t = \mu_t + \beta_t X_t + S_t + \varepsilon_t \ ,$$

where Y_t is the observed unemployment rate at time t; μ_t is a time varying intercept or trend term; β_t is a (1xk) row vector of time-varying coefficients; X_t is a column vector of explanatory (regressor) variables at time t; S_t is a seasonal component; and ε_t is an error term.

The time-varying trend is represented by a locally smooth linear trend with a random level, μ_t, and slope, $\gamma_{t.}$. This is expressed as

$$\mu_t = \mu_{t-1} + \gamma_{t-1} + u_t$$

$$\gamma_t = \gamma_{t-1} + u^*_t,$$

where u_t and u^*_t are independent white noise terms with zero expectations and variances σ^2_u and $\sigma^2_{u^*}$. Similarly, the regression coefficient vector is described by

$$\beta_t = \beta_{t-1} + \xi_t \ ,$$

where ξ_t is a white noise vector with zero expectation and covariance matrix Diag($\sigma^2_{\xi_1},..,\sigma^2_{\xi_k}$).

The seasonal component is modeled by

$$S_t = \sum_{j=1}^{6} S_{jt} \ ,$$

where

$$S_{jt} = \cos(\omega_j)S_{j,t-1} + \sin(\omega_j)S^*_{j,t-1} + \zeta_{s_j} \ ,$$

$$S^*_{jt} = -\sin(\omega_j)S_{j,t-1} + \cos(\omega_j)S^*_{j,t-1} + \zeta^*_{s_j} \ ,$$

$$\omega_j = 2\pi p_j^{-1}, \ \ p = \{12, 6, 4, 3, 2.4, 2\},$$

and the ζ_{s_j} and $\zeta^*_{s_j}$ are independent white noise disturbances with zero means and constant variance.

This model provides, as a special case, the standard time series regression model in which the intercept and regression coefficients are fixed and the seasonal component is represented by monthly dummy variables. The role of the time varying intercept is to capture long run variation in the unemployment rate that is not reflected in the explanatory variables. Similarly, the purpose of the seasonal component is to account for seasonal movements in the rate that can not be fully explained by the regressor variables. Seasonal movements are allowed to slowly change in magnitude over the sample period. The error term, ε_t, accounts for survey measurement error and can be modeled by a stationary process. For national data, the relative variance of the CPS survey errors are small enough so that the autocorrelation structure can be ignored; thus we let ε_t be a zero mean, white noise process with variance $\mathrm{s}_{\mathrm{e}}^{2}$. Additionally, we assume that all disturbance terms in the model are normally distributed.

In our application, we used two explanatory variables in the model. These are CESEM_t , employment estimated by the Current Employment Statistics survey; UI $_t$, worker claims for unemployment insurance benefits; and POP_t, representing the civilian noninstitutional population, 16 years and over. These variables are defined below.

$$\mathrm{CESEP}_t = 100(\mathrm{CESEM}_t/\mathrm{POP}_t)$$

$$\mathrm{CLR}_t = 100(\mathrm{UI}_t/\mathrm{CESEM}_t)$$

Two models were fitted. One model includes both variables, and the other includes only the CLR variable. For parameter estimation and signal extraction, the models were expressed in state space form. The parameters were estimated based on a maximum likelihood procedure, using the Kalman filter to estimate the likelihood function. Given the parameters of the system, the Kalman filter was used to optimally decompose a sample observation into its signal and noise components[5].

Regression models with autoregressive disturbances. In this approach, multiple regression models with autocorrelated errors are formulated as follows[6].

$$Y_t = \mathrm{b}X_t + \mathrm{e}_t$$

$$\mathrm{e}_t = \sum_{i=1}^{p} \mathrm{q}_i \mathrm{e}_{t-i} + \mathrm{u}_t$$

These models were estimated in both level and difference form. Variables used in the model are as follows:

RMPY7 = $(\mathrm{POP}_t)^{-1}$Y7 - 3559$(\mathrm{POP}_t)^{-1}$
RMPY6 = $(\mathrm{POP}_t)^{-1}$Y6 - 50416$(\mathrm{POP}_t)^{-1}$
Y7 = Unemployment insurance claims
Y6 = CES employment
POP = Civilian population 16+
CPSUER = CPS unemployment rate
DR = First difference of CPS unemployment rate
RMDIF = Lag(RMPY7 - .5RMPY6)
MMDINT = [RMPY7 - 0.46RMPY6 + 0.43RMDIF][$\overline{POP}$]$^{-1}$[POP - $\overline{POP}$]
$\overline{POP}$ = mean of POP
TREND = $\left[1+e^{\{0.375(t-122)\}}\right]^{-1}$

The level model contained RMPY7, RMPY6, RMDIF, MDINT, TREND, 11 seasonal dummies, and 11 cross-product terms of TREND with the seasonal dummies. The difference model contained DRMPY7, DRMPY6, DRMDIF, 11 seasonal dummies, and the 11 season-by-TREND cross product terms. The notable features of these regression models were the inclusion of the TREND variable, estimated from fitting a logistic function to the unadjusted unemployment rate and the detrending of several variables. An additional variable, the proportion of teenagers 16 and older in the civilian noninstitutional population was included in the difference model to account for changing seasonality in May and June. Other auxiliary variables were tried in the regression model, but did not improve the model fit; these variables included the help-wanted index and number of hours worked in the manufacturing sector.

ARIMA models

ARIMA modeling is one of the most frequently used approaches to short term forecasting[7]. These models allow for a wide variety of potential forecast functions for extrapolating a time series from its own past. However, because forecasts are required for up to the first 5 months of 1994, the ARIMA univariate model has limited application, as the forecasts standard errors for this type of model increase considerably as the forecast period is extended.

With multi-period forecasting, models that use related series (when the values of those series are available during the forecast period) are preferred because their forecast errors are likely to be smaller. However, ARIMA models are useful benchmarks for comparison, because they often produce high quality forecasts over the first few periods of the forecast range. ARIMA models that were useful for one-step-ahead forecasts were the $(3,1,0)(0,1,1)_{12}$ and the $(0,2,2)(0,1,1)_{12}$.

Endnotes

1. D. L. Kostanich, and L. S. Cahoon, "Effect of Design Differences Between the Parallel Survey and New CPS," *CPS Bridge Team Technical Report 3,* (Bureau of Labor Statistics, 1994).

2. A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter (*Cambridge, Cambridge University Press, 1989).

3. W. A. Fuller, *Introduction to Statistical Time Series* (New York, Wiley, 1976).

4. G. E. P. Box, and G. M. Jenkins, *Time Series Analysis: Forecasting and Control* (San Francisco, Holden-Day, 1976).

Table 2. Official and Predicted Unemployment Rates Based on the old CPS Design January 1993 - October 1994

			Seasonally adjusted		
	Not seasonally	Standard	Unemployment	90% Confidence interval	
Month	adjusted	error*	rate	lower	upper
Official					
January 93	7.9	0.12	7.1	6.9	7.3
February	7.7	0.12	7.0	6.8	7.2
March	7.3	0.12	7.0	6.8	7.2
April	6.8	0.11	7.0	6.8	7.2
May	6.7	0.11	6.9	6.8	7.1
June	7.1	0.11	6.9	6.8	7.1
July	6.9	0.11	6.8	6.6	7.0
August	6.5	0.11	6.7	6.6	6.9
September	6.4	0.11	6.7	6.5	6.9
October	6.3	0.11	6.7	6.5	6.9
November	6.1	0.10	6.5	6.3	6.6
December	6.0	0.10	6.4	6.2	6.6
Predicted					
January 94	6.9	0.17	6.3	6.0	6.6
February	7.0	0.20	6.4	6.1	6.7
March	6.6	0.22	6.3	5.9	6.7
April	5.9	0.24	6.1	5.7	6.5
May	6.0	0.26	6.1	5.7	6.5
June	6.1	0.28	5.9	5.4	6.3
July	6.1	0.29	6.0	5.5	6.4
August	5.8	0.31	6.0	5.5	6.5
September	5.6	0.32	5.9	5.4	6.5
October	5.5	0.34	5.9	5.3	6.4

*Standard errors are based on rates that are not seasonally adjusted and are used to construct the confidence intervals.