MEASURING EMPLOYMENT FROM BIRTHS AND DEATHS IN THE CURRENT EMPLOYMENT STATISTICS SURVEY

Diem-Tran Kratzke, Heidi Shierholz, Steve Woodruff, Bureau of Labor Statistics.
Steve Woodruff, BLS, 2 Massachusetts Ave, N.E. Washington, DC 20212

KEY WORDS: Seemingly Unrelated Regression, Generalized Least Squares, Mixed Estimation.

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics. The authors thank Alan Dorfman of the Bureau of Labor Statistics for his review and comments.

1. Overview of the Current Employment Statistics Survey

The Current Employment Statistics (CES) survey is a nationwide monthly payroll survey of close to 400,000 business establishments. It provides current estimates of employment, hours, and earnings in industry and area detail for the 50 states, the District of Columbia, Puerto Rico, and the Virgin Islands. The Bureau of Labor Statistics (BLS) is currently redesigning the CES survey. The research presented here is one component of the redesign.

The CES sampling frame is called the Business Establishment List (BEL). BEL is a list of seven million establishments whose primary source is the quarterly contribution reports that employers file with their state Unemployment Insurance (UI) agencies. Quarterly BEL data are available to statisticians and economists at BLS approximately nine months after the end of a quarter.

The sampling frame is updated yearly to account for changes such as new businesses starting up (births) and existing businesses going out-of-business (deaths). Births that occur in between frame updates can be problematic; it is very difficult and expensive to maintain a comprehensive and timely sampling frame for new births. New deaths also pose a problem; though it should be possible to directly measure them from the sample, in practice it can be very hard to distinguish between deaths and nonrespondents because of the vast sample size and short data collection period of the CES.

This paper focuses on the cost effective and reliable measurement of the gain of employment from births that have not yet been added to the frame (birth employment) and the loss of employment from deaths that have not yet been removed from the frame (death employment).

2. Research Approach

The continuous portion of the universe consists of all units that are not births, and their employment is called continuous employment. The total employment figure that the CES survey measures is continuous employment plus birth employment. Our approach is not to measure the birth employment directly, but rather to predict the net employment, which is the difference between birth employment and death employment. Since birth employment is equal to death employment plus net employment, it follows that total employment is equal to the sum of continuous employment, death employment, and net employment. In estimating total employment, continuous employment and death employment are both estimated from the CES sample. Death employment is estimated by retaining all dead UI accounts in the sample and imputing their employment with the same imputing technique used for non-respondents. Net employment is predicted using a statistical model.

Operationally, this approach is advantageous; a sample unit is imputed whether it is a non-respondent or a death, so it is unnecessary to distinguish between the two.

This paper discusses methods for modeling net employment and examines the ability of net models to predict net employment.

3. Data

Four historical CES data frames were constructed for a simulation study of the CES redesign. These frames were used for the research of this paper. They are called Frame90, Frame91, Frame92, and Frame93. Frame(yy) consists of all UI accounts that existed at March '(yy-1) plus all new UI accounts that occurred between April '(yy-1) and March '(yy).

The employment estimates from April of one year to September of the next year constitute a set of estimates. Four sets of estimates that correspond to the four data frames mentioned above were produced: April '90 to September '91; April '91 to September '92; April '92 to September '93; and April '93 to September '94. For a given estimation period, the estimates do not reflect any gain in employment due to new businesses that have occurred since the beginning of that period. For example, the May '90 estimate does not reflect any gain in employment due to businesses that were born in April '90 or May '90.

The above file structure reflects the fact that the CES frame has historically been updated when the first quarter (January through March) of BEL data becomes available each year. In this study, we assume that the data becomes available in October. At this time, revised estimates are computed for April through September, using an updated sample. The sample that was used for the first round of estimates is updated to include a sample of births and to exclude any deaths that have occurred since April of previous year until March of the current year. In addition, the March population employment is used to ratio adjust the revised estimates in a process called benchmarking. March is referred to as the benchmark month and the employment in March is called the benchmark employment. Benchmarking from the March employment is also used for estimates from October of the current year through September of the following year, at which time a new benchmark employment will be available.

In order to simulate the net employment, four files of births and four files of deaths were constructed for the same periods that estimates were produced. That is, file 1 consists of births (or deaths) that occur between April '90 and September '91, file 2 consists of births (or deaths) that occur between April '91 and September '92, and so forth.

A birth file consists of UI accounts that have zero employment in the twelve months preceding the file period, and at least one month of positive employment in the file period. Birth dates were assigned to each birth unit as the first month with positive employment.

A death file consists of UI accounts that have at least one month of positive employment in the twelve months preceding the file period, zero employment for the last month of the file period and zero employment for the three months following the file period. The requirement of an additional three months of zero employment is given in order to minimize the chance of defining seasonal businesses that may have zero employment for some months of the year as deaths. Death dates were assigned to each death unit as the first month of lasting zero employment.

In our study, data from the first two files of births and deaths were used to model the net employment, and the parameters from the models are used to predict net employment for the last twelve months of the last file. This timing reflects the availability of data for modeling during the actual CES production.

4. Model of Net Employment

Net employment is calculated as the difference between cumulative birth employment and cumulative death employment. At any month on a file, cumulative birth employment is defined as the sum of employment at that month of all births since the benchmark month, and cumulative death employment is defined as the sum of the last positive employment of all deaths since the benchmark month. Net employment at month t is represented by N_t, where $t = 0$ at the benchmark month.

a. Explanatory Variables

Two explanatory variables are used in the model. The first is the change in the industry employment level since the benchmark. For month t, it is defined as the difference between the employment at month t as estimated by the sample and the benchmark employment. It is denoted by Z_t. The second explanatory variable is the net employment for the previous year in the same month, or N_t^p. This variable makes use of the fact that there is a high correlation between employment in the same month from one year to the next.

b. Seemingly Unrelated Regression

When applied to CES data, classical normal linear regression models do not incorporate all available knowledge about the regression equations and the variables involved. They ignore the fact that errors in a regression equation for an industry are correlated with the errors in the regression equation for the total (all industries combined). A system of equations that has a property of this type is referred to as a system of seemingly unrelated regression (SUR) equations. Generalized least squares (GLS) parameter estimates for a system of SUR equations can be derived that take into account the mutually correlated regression errors in different equations. In what follows, such GLS parameter estimates are derived using Z_t and N_t^p as independent variables. The application to any combination of independent variables is straightforward.

The net predictor described below is based on partitioning a set or aggregate (AG) of business establishments into sub-aggregates (SAGs) and allowing the aggregate level data to lend strength to the sub-aggregates. The aggregate used for the research is the national total and the sub-aggregates the industry divisions. When industry information is available, establishments are assigned to one of the eight industries: Mining, Construction, Manufacturing, Wholesale, Retail, Transportation and Public Utilities (TPU), Finance Insurance and Real Estates (FIRE), and Services. Otherwise, they are assigned to the unclassified industry. In the study, a SAG refers to a classified industry at the national level and the AG is the national total. Here, the unclassified industry will not be included in the SUR system, even though a linear regression could be applied to these establishments.

Let the subscript j refer to a SAG , ($1 \le j \le 8$), u refer to the unclassified industry, and s refer to the AG. Then $N_{ts} = \sum_{j=1}^{8} N_{tj} + N_{tu}$ is the national total net employment at time t.

Let $X_{tj} = \left(Z_{tj}, N_{tj}^p\right)$and $\beta_j' = \left(\alpha_{1j}, \quad \alpha_{2j}\right)$. For SAG j, the linear model then takes the form

$$N_{tj} = X_{tj}\beta_j + \varepsilon_{tj} \quad \text{for} \quad 1 \le j \le 8$$

with $E(\varepsilon_{tj}) = 0$ and $Var(\varepsilon_{tj}) = a_t B_j \sigma^2$

and for the unclassified industry, we have

$$N_{tu} = X_{tu}\beta_u + \varepsilon_{tu}$$

with $E(\varepsilon_{tu}) = 0$ and $Var(\varepsilon_{tu}) = a_t B_u \sigma^2$.

For fixed *t*, the $\{\varepsilon_{t1}, \varepsilon_{t2}, ..., \varepsilon_{tw}, \varepsilon_{tu}\}$ are uncorrelated, *B* denotes the benchmark employment, a_t is an increasing function of *t* (to be further discussed below), and $t = 1,2,...,T$ (note that, for this problem, *T* = *18*).

We have $N_{ts} = \sum_{j=1}^{8} N_{tj} + N_{tu}$ and $X_{ts} = \sum_{j=1}^{8} X_{tj} + X_{tu}$ which imply

$$N_{ts} = \sum_{j=1}^{8} X_{tj}\beta_j + X_{tu}\beta_u + (\sum_{j=1}^{8} \varepsilon_{tj} + \varepsilon_{tu}).$$

Following the above linear regression form, let N_{ts} also be given by $N_{ts} = X_{ts}\beta_s + \varepsilon_{ts}$. Then by equating these two expressions for N_{ts}, we have:

$$\varepsilon_{ts} = \sum_{j=1}^{8} X_{tj}(\beta_j - \beta_s) + X_{tu}(\beta_u - \beta_s) + (\sum_{j=1}^{8} \varepsilon_{tj} + \varepsilon_{tu})$$

and this implies

$$E(\varepsilon_{ts}) = \sum_{j=1}^{8} X_{tj}(\beta_j - \beta_s) + X_{tu}(\beta_u - \beta_s)$$

and $$Var(\varepsilon_{ts}) = \sigma^2 a_t (\sum_{j=1}^{8} B_j + B_u) = \sigma^2 a_t B_s.$$

It is reasonable to assume that $E(\varepsilon_{ts})$ is negligible compared to $X_{ts}\beta_s$. It is therefore set to zero in our models.

The above equations can be written as:

$$N_{.j} = X_{.j}\beta_j + \varepsilon_{.j},$$

and $$N_{.s} = X_{.s}\beta_s + \varepsilon_{.s},$$

where $N_{.j}$ and $N_{.s}$ are $(T \times 1)$ vectors, $X_{.j}$ and $X_{.s}$ are $(T \times 2)$ matrices, β_j and β_s are (2×1) vectors, and $\varepsilon_{.j}$ and $\varepsilon_{.s}$ are $(T \times 1)$ vectors.

Let $A = \{a_{tt'}\} = \{Cov(N_{tj}, N_{t'j})\}$. Then A is a $(T \times T)$ matrix with the following form:

$$A = \begin{pmatrix} a_1 & a_1 & a_1 & . & . & . & a_1 \\ a_1 & a_2 & a_2 & a_2 & . & . & a_2 \\ a_1 & a_2 & a_3 & a_3 & a_3 & . & a_3 \\ a_1 & a_2 & a_3 & a_4 & a_4 & . & a_4 \\ . & . & a_3 & a_4 & . & . & . \\ . & . & . & . & . & a_{T-1} & a_{T-1} \\ a_1 & a_2 & a_3 & a_4 & . & a_{T-1} & a_T \end{pmatrix}.$$

Given *A*, let $E(\varepsilon_{.s}\varepsilon'_{.s}) = \sigma^2 B_s A$, and

$$E(\varepsilon_{.j}\varepsilon'_{.i}) = \begin{cases} \sigma^2 B_j A, & \text{when } j = i \\ 0, & \text{when } j \neq i \end{cases}.$$

To incorporate the correlation of regression errors between SAGs within the AG and the whole AG ($E(\varepsilon_{.s}\varepsilon'_{.j}) = \sigma^2 B_j A$), the system of SUR equations is compressed into one equation. In the following matrix expression, 0 denotes a $(T \times 2)$ matrix of zeros.

$$\begin{pmatrix} N_{.s} \\ N_{.1} \\ N_{.2} \\ N_{.4} \\ \vdots \\ N_{.8} \end{pmatrix} = \begin{pmatrix} X_{.s} & 0 & 0 & 0 & 0 & 0 \\ 0 & X_{.1} & 0 & 0 & 0 & 0 \\ 0 & 0 & X_{.2} & 0 & 0 & 0 \\ 0 & 0 & 0 & X_{.3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & X_{.8} \end{pmatrix} \begin{pmatrix} \beta_s \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_{.s} \\ \varepsilon_{.1} \\ \varepsilon_{.2} \\ \varepsilon_{.3} \\ \vdots \\ \varepsilon_{.8} \end{pmatrix}.$$

Let the compact version of this expression be

$$\text{N} = \text{X}\beta + \varepsilon,$$

where N, β, X, and ε are the associated matrices in the above system of equation.

The covariance matrix of ε is:

$$\Sigma = \sigma^2 \begin{pmatrix} B_s & B_1 & B_2 & B_3 & . & . & .. & B_8 \\ B_1 & B_1 & 0 & 0 & 0 & . & .. & 0 \\ B_2 & 0 & B_2 & 0 & 0 & . & .. & 0 \\ B_3 & 0 & 0 & B_3 & 0 & . & .. & 0 \\ . & 0 & 0 & 0 & . & . & .. & 0 \\ . & . & . & . & . & . & 0. & 0 \\ . & . & . & . & . & 0 & B_7 & 0 \\ B_w & 0 & 0 & 0 & 0 & 0 & 0 & B_w \end{pmatrix} \otimes A,$$

where ⊗ is the Kronecker product.

The least squares estimate of β is $\hat{\beta} = (X\Sigma^{-1}X)^{-1} X\Sigma^{-1}N$ (4b.1).

c. Mixed Estimation

Since we are only interested in estimating β in order to use it to predict future net employment, it proved useful to apply a regression technique called mixed estimation from Belsley, Kuh, and Welsch (1980). The effect of this additional structure is to place stochastic constraints on the components of β so that the coefficient of the Z_{tj} is given extra importance. Recall that Z_{tj} estimates the change in employment since the benchmark month. This variable is thus an indicator of current economic movement, which we want the regression equation to capture. It is given extra emphasis in hopes that it will contribute to more accurate employment estimates in times of an unsually strong or weak economy.

Let the stochastic constraint that we wish to force on β be given by the linear model c=Rβ + ξ where E(ξ)=0 and Var(ξ)=Σ_1. R is a matrix with the same number of columns as β and each row is a constraint on β. In our case, it is logical to constrain each pair of components in the $\{\beta_s, \beta_1, \ldots.\beta_8\}$ separately. Thus R has 9 rows with each row being a constraint on the two components of a vector of regression coefficients from this set. The quantity c is a 9x1 vector of constants, and ξ is a random vector that is uncorrelated with ε.

The estimation proceeds by augmenting N, X, and ε to give the constrained linear model used to estimate β:

$$\begin{pmatrix} N \\ c \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix}\beta + \begin{pmatrix} \varepsilon \\ \xi \end{pmatrix} \text{, where } Var\begin{pmatrix} \varepsilon \\ \xi \end{pmatrix} = \begin{pmatrix} \Sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix}.$$

The mixed estimate of β is

$$\hat{\beta}_M = (X\Sigma^{-1}X + R'\Sigma_1^{-1}R)^{-1}(X\Sigma^{-1}N + R'\Sigma_1^{-1}c) \qquad (4c.1).$$

The following discussion gives the reason for the choice of R, c, and Σ_1. If we regress the Z_{tj} on the net employment, the regression coefficient is roughly around .15. If we regress N_t^P on the net employment, the coefficient is roughly 1. Thus both $(.15)Z_{tj}$ and N_{tj}^P estimate N_{tj}. If we were to take a weighted average of these two estimates -- as in composite estimation -- then with the right weights we would have an improved estimate of N_{tj}. Write this composite as: $(1-a)(.15)Z_{tj} + aN_{tj}^P$. Then let *b*=(1-*a*)(.15). If *a* is chosen optimally then *b* and *a* should roughly equal α_{j1} and α_{j2} (the components of β_j) respectively. The result is the restriction $\alpha_{j1} + (.15)\alpha_{j2}$ =.15 for all *j*.

Thus in the equation c=Rβ + ξ, we have R=I⊗(1,.15) where I is the 9x9 identity matrix and ⊗ is the Kronecker product, c is the column vector with .15 in each position, and Σ_1 is the diagonal matrix with .00004 along the diagonal. The covariance matrix Σ_1 forces a relatively small variation from the purely deterministic constraint, c=Rβ.

d. Parameter Estimation

The empirical results following the stochastic description of the process that generates the net employment showed that it is appropriate to set $a_t = \frac{t(t+1)}{2}$. The vector of regression coefficients is then only a function of only one unknown parameter, σ^2. Initially set $\hat{\sigma}^2 = 1$ and estimate β according to equation (4b.1) or (4c.1). This initial estimate of β, $\ddot{\beta}$, is then used to obtain improved estimates of σ^2, which are estimated in the "usual" way from the squared residuals.

Thus the new estimate of $\hat{\sigma}^2$ is $\hat{\sigma}^2 = \left(\frac{1}{18x9}\right)\left(\sum_{j,t}\left(\frac{N_{jt} - X_{jt}\ddot{\beta}_j}{\sqrt{B_j a_t}}\right)^2\right)$, where the summation is over the SAGs and the 18 months of the modeling file and $\ddot{\beta}_j$ is the j^{th} SAG component of $\ddot{\beta}$.

5. Empirical Studies

Recall that models are built using data from the first two birth and death files. Four different regression models are presented here. Each model regresses the net employment on the change in employment since benchmark month, Z_t, and the previous year's net employment, N_t^P. The four different models are:

a. *Ordinary Least Square* (OLS) – The regression is run separately for each industry and for the national total, with uncorrelated errors that have zero expectation and constant variance.

b. *Seemingly Unrelated Regression 1* (SUR1) – This model follows the variance structure presented in Section 4b, with the exception that the off-diagnonal elements of Σ are all zeros, indicating that $E(\varepsilon_s \varepsilon_j') = 0$.

c. *Seemingly Unrelated Regression 2* (SUR2) – This model follows the SUR system as described in Section 4b.

d. *Seemingly Unrelated Regression 3* (SUR3) – This model follows the SUR system as described in Section 4b along with the mixed estimation described in Section 4c.

Net employment was predicted for the last twelve months of file 4. The prediction error is defined as the difference between the true net employment and the predicted net employment. The absolute relative prediction error (ARPE) averaged over the twelve months was computed as:

$$\overline{ARPE} = \frac{1}{12}\sum_{t=1}^{12}\left|\frac{N_t - \hat{N}_t}{N_t}\right|.$$

The absolute relative prediction error was also computed for the month of March, which is of particular concern because it is the benchmark month.

$$ARPE(March) = \left|\frac{N_6 - \hat{N}_6}{N_6}\right|.$$

In the above equations, t=1,2,…,12 correspond to the months October '93 through September '94.

The smaller the value for $\overline{ARPE}$ and $ARPE(March)$, the better the model is predicting the net employment. Values of $\overline{ARPE}$ and $ARPE(March)$ that are greater than one indicate that using the model is actually detrimental. In these cases, the estimate of total employment would be more accurate if the sample estimate alone were used.

6. Results

The following tables show $\overline{ARPE}$ and $ARPE(March)$ for the four models at the total and industry level.

Table I

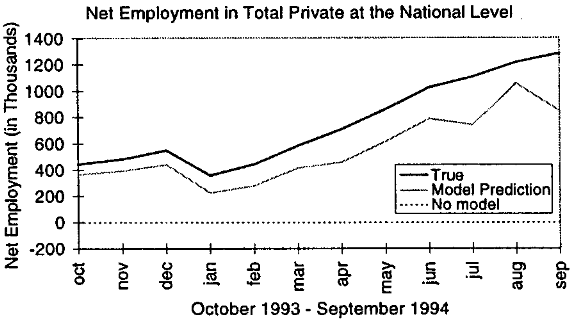
	$\overline{ARPE}$			
	OLS	SUR1	SUR2	SUR3
Total Private	0.273	0.455	0.570	0.272
Mining	1.046	1.382	3.031	3.686
Construction	0.510	1.226	0.809	0.579
Manufacturing	10.125	9.300	1.697	1.901
TPU	0.732	0.809	0.419	0.811
Wholesale	5.015	1.147	1.677	3.125
Retail	0.124	0.110	0.530	0.232
FIRE	2.307	2.687	1.139	1.341
Services	0.181	0.245	0.539	0.240

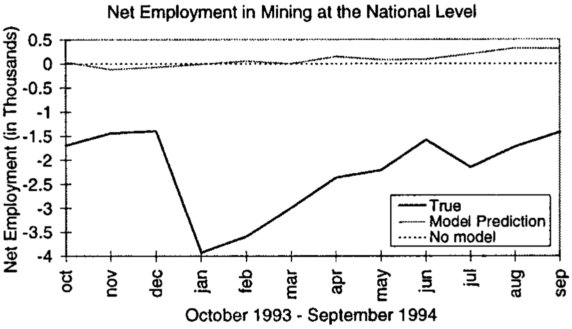
Table II

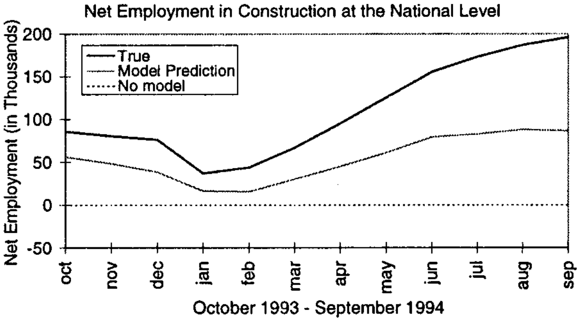
	$ARPE(March)$			
	OLS	SUR1	SUR2	SUR3
Total Private	0.292	0.463	0.568	0.263
Mining	1.001	1.200	1.888	2.141
Construction	0.546	1.531	0.909	0.712
Manufacturing	71.842	67.109	11.358	13.503
TPU	0.647	0.862	0.113	0.752
Wholesale	1.880	0.483	0.162	0.296
Retail	0.087	0.067	0.561	0.254
FIRE	2.189	2.574	1.082	1.231
Services	0.092	0.162	0.483	0.146

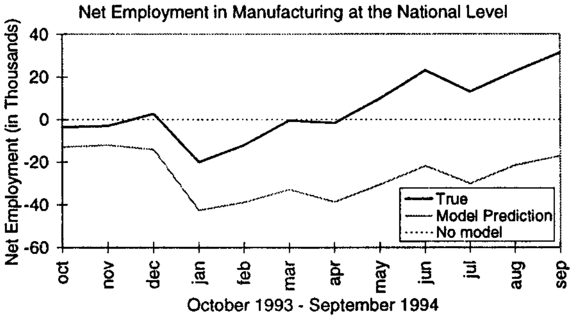
From the two tables, we see that for Mining, Manufacturing, and FIRE, none of the models was helpful in predicting the net employment. For Wholesale Trade, the three SUR models were helpful in March, but not on average. For the four remaining industries, OLS is the best model for Construction and Services, SUR1 is the best model for Retail Trade, and SUR2 is the best model for TPU. The model SUR3 is best for Total Private only. In three out of the four industries where modeling is helpful, using the mixed estimation technique along with the seemingly unrelated regression is better then using the seemingly unrelated regression alone. However, the simplest model, OLS, is the best or second best model for all the industries where modeling is helpful.

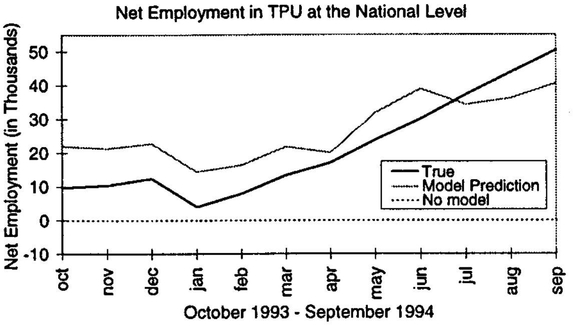
The following plots show the true net employment and the net employment as predicted by the OLS model. The graphs also show a reference line at zero, which demonstrates whether or not the model is helpful by comparing it to using no model at all (or equivalently, by predicting that the net employment is zero, and therefore, assuming that birth and death employment are roughly equal).

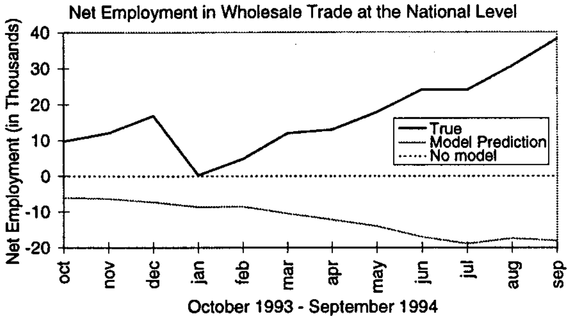


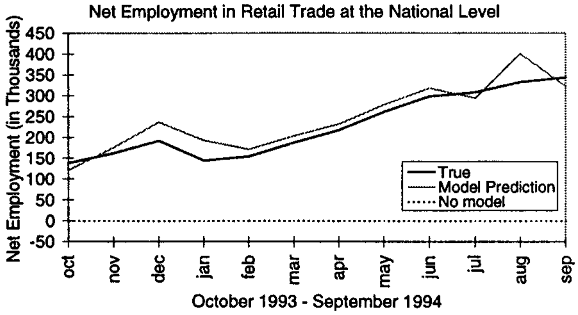


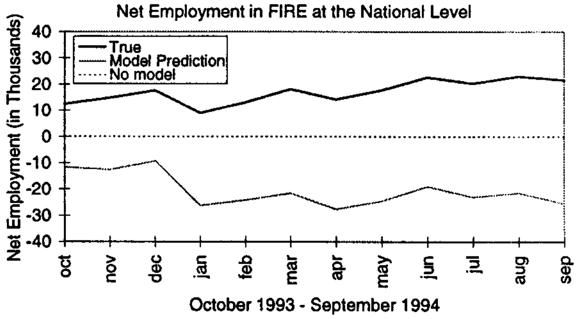


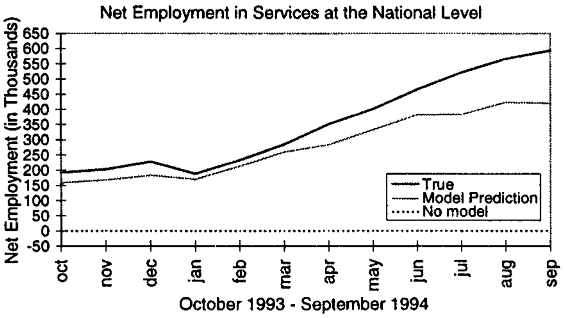












7. Conclusions

In the major industry divisions of Mining, Manufacturing, Wholesale Trade, and FIRE, it was found that a net model should not be used, that instead the sample estimate alone should be used.

In Construction, TPU, Retail Trade, and Services, it was found that the OLS model with employment since the benchmark month and the previous year's net employment as explanatory variables should be used to predict net employment. Though it is not the best model in all the industries, the OLS model outperforms the SUR models across industries. The OLS model also has the advantage that it is simpler in concept and is easier to implement.

8. Future Research

The modeling for this paper was conducted at the national level. Modeling at the state level is a logical extension.

The research presented here was conducted using simulated data. The next step in evaluating the net modeling approach will be to test it in production in order to make an assessment of how the approach works under the real conditions of the CES.

The behavior of the net employment in all industries should be monitored over time in order to verify or revise the conclusion that modeling will be beneficial only in Construction, TPU, Retail Trade, and Services.

9. References

Shail B., Stamas G., Brick M., (1997), "Sample Redesign for the Current Employment Statistics Survey," *ASA Proceedings of the Survey Research Methods Section.*

Belsley D. A., Kuh E., Welsch R. E. (1980), *Regression Diagnostics,* John Wiley & Sons, Inc.

Kmenta, J. (1971), *Elements of Econometrics,* The Macmillan Company, New York.

Rao C. R. (1973*), Linear Statistical Inference and Its Applications*, John Wiley & Sons.

Werking, G. (1997), "Overview of the CES Redesign," *ASA Proceedings of the Survey Research Methods Section.*