AN APPLICATION OF REGRESSION SUPERPOPULATION MODELS IN THE CURRENT EMPLOYMENT STATISTICS SURVEY

Christopher H. Johnson and Stephen M. Woodruff, Bureau of Labor Statistics
Christopher H. Johnson, 441 G St. NW, Washington, DC 20212

KEY WORDS: model–based estimation, nonignorable nonresponse

1. Introduction

The Bureau of Labor Statistics' (BLS) Current Employment Statistics (CES) Survey gathers data monthly from over 340,000 nonagricultural business establishments for the purpose of estimating total employment, women and production workers, hours, and earnings. Estimates are made for over 1500 industry cells, complementing the demographic detail provided by estimates of employment from the Current Population Survey. Current monthly estimates of employment level and month to month change in employment are of primary importance to the users of these data. In addition to the CES survey, each state conducts a complete count of the employment of its business population every quarter following the guidelines of the Unemployment Insurance (UI) system. Except for a few industries exempt from UI coverage, this complete count is used by the CES as a benchmark to which survey estimates are revised and to which they are compared to derive a measure of error.

This paper compares the current estimator, called the link relative (LR) estimator, and a proposed estimator called the Link 90 (L90) for estimating employment level. The sample data collected in the CES survey, often subject to considerable nonresponse, are used in this empirical comparison.

The sample of business establishments used in this survey is substantially fixed over time and is composed of most large establishments, with a less extensive sample of smaller establishments. The CES sample is obtained by soliciting businesses establishments until a "sufficient" number agree to participate, and thus no sampling distribution (or response mechanism) can be assumed. Variability is largely due to nonresponse, and in the simulation studies to be described later, the sample will be fixed and the response mechanism will be the sole source of variability between replicates.

The CES survey and the LR estimator have been studied in depth over the last two decades by Royall (1981), Madow and Madow (1978), West (1982, 1983), and Woodruff (1982, 1983, 1989). Newer methodologies which take advantage of advances in data processing capabilities are being developed for this survey. The proposed L90 is one such estimator, which would have been infeasible four or five decades ago when the LR was adopted for the CES survey.

Both the LR estimator and the L90 make use of known relationships between employment during adjacent months in a business establishment, but the L90 makes fuller use of available sample data: To be included in the link between two adjacent months, a unit must have responded for these months by the time that link is computed. Thus, sample units which have data for only one of these two months will have their data excluded from the LR estimator, making it inefficient for estimating the most current months' employment. The L90 solves this apparent inefficiency by making direct use of the strengths of the LR estimator and designing around its weaknesses. Both estimators use the well–documented [West (1982, 1983), Madow and Madow (1978)] relationship that the conditional expected value of employment in month k for an establishment, given its employment in month k–1, is proportional to its employment in month k–1.

Considerable research has been done into the LR estimator and, in spite of certain problems with it that have been pointed out by most of the above authors, nothing sufficiently better to justify large–scale operational changes has been found to replace it. This research may have similar results, but preliminary testing suggests that the proposed L90 merits serious consideration.

Section two describes the flow of CES sample data into the BLS and how they are then used to construct a series of estimates of total employment for a given month. Section three describes and derives the L90. Section four describes a simulation study to compare the two estimators.

2. CES Data Flow

The CES survey is used to estimate employment level and the change in employment each month. Establishments report data for the pay period that includes the 12^{th} of the month. Employment level is determined once a year from the complete count, or benchmark, of businesses conducted by the states' unemployment insurance system.

An employment link for a basic estimating cell (a combination of industry and size categories) for month j is the quotient of total employees for month j in the current matched sample (all sample units that have data for both month j and the previous month, j–1, and that have passed certain edits) and total employees for month j–1 in this matched sample. Denote this link $\hat{\beta}_j$. The LR estimate LR_j of employment for a basic cell for month j is the product of this link and the LR estimate for the previous month, LR_{j-1}. Thus $LR_j = \hat{\beta}_j LR_{j-1}$. By convention, we let j=0 denote the benchmark month and LR_0 the benchmark employment for the cell.

The first estimates of employment computed for the current month are preliminary figures based on the initially available microdata in the current matched sample. These are called first closing estimates, based on the 50% or so of sample reports received by the closing date for the current month. Before the preliminary employment estimates for the current month are computed, more complete matched sample totals for the previous month are obtained; that is, reports of last month's data not received in time to be included in last month's first closing estimates are added. These augmented matched sample totals are used in recomputing both links and LR estimates for the previous month and determining the revised estimates for that month. This link uses data from about 70% of the sample units. These second closing estimates for the previous month are then multiplied by the current link to obtain first closing estimates for the current month. The extension to third closing estimates is immediate: These incorporate data not received by the cut off date for second closing estimates and are based on data from about 90% of the sample reports. Given this piecemeal (over time) arrival of sample data, it would be appropriate to place a second subscript, t, on $\hat{\beta}_j$ and LR_j. For example, $\hat{\beta}_{jt}$ would denote the link for month j computed from all matched data available at time t. To minimize notational clutter, we omit this subscript, trusting that the value of t will be clear from context.

Note that the LR estimator for month m uses only sample data for month m and earlier months and ignores much of the nonmatched data. The proposed L90 uses all of the sample data which pass the edits, and for second and third closing estimates for month m the L90 uses not only past and current CES data but also data from months m+1 and m+2.

3. Model and Estimator

We use two types of information to design an estimator. The first type of information is numerical data, which is expressed as a row vector, Y_i, attached to each unit i in the population of N units. This row vector consists of two subvectors, A_i and T_i. The first subvector, A_i, consists of the auxiliary variables, often defined as those items known for all units in the population. The second, T_i, consists of the target variables, the finite population means of which we wish to estimate. These targets are observed for only a sample of the population units. Because of nonresponse, a subset of the targets is missing for each sample unit, so the subset of observed targets varies from sample unit to sample unit. In our CES application, A_i is a scalar and T_i is a $1 \times l$ vector.

The second type of information relevant to estimator design is knowledge about the relation between the above types of data and the population characteristic we wish to estimate. This information is usually expressed through modeling. The models may include relationships between the study characteristic and the method by which data were either observed or missing for a population unit, relationships often expressed as sampling distributions or response mechanisms. Another useful type of data relationship is the one between the data items themselves. Such relationships can often be adequately described in terms of the covariances between pairs of random variables representing pairs of different data items (auxiliary and target variables). These covariances are derived from superpopulation models.

We can represent the observed (responding) targets and all the auxiliary variables for the i^{th} population unit as $Y_i = (A_i, T_i\, \chi_i)$, where χ_i is the $l \times l_{Ri}$ response indicator matrix for the $1 \times l$ row vector of target variables T_i attached to the i^{th} sample unit. The term l_{Ri} is the number of variables in Y_i for which there are responses; necessarily, $l_{Ri} \le l$. χ_i is constructed from the identity matrix of order l by deleting each column j of this identity matrix for which target variable j in T_i is a nonresponse. Thus Y_i contains the i^{th} unit's auxiliary variable(s) and only those target variables we observed for the i^{th} unit. We let $\mu = (\mu_A, \mu_T)$ be the vector of auxiliary and target means. The mean of the auxiliary variables, μ_A, is known. The mean of the target variables, μ_T, is to be estimated. Letting I be the identity matrix of order equal to the number of auxiliary variables (components of an A_i) and $X_i = \begin{bmatrix} I & 0 \\ 0 & \chi_i \end{bmatrix}$, we can also write

$$Y_i = (A_i, T_i\, \chi_i) = (A_i, T_i) \begin{bmatrix} I & 0 \\ 0 & \chi_i \end{bmatrix} = (A_i, T_i)\, X_i.$$

If we add in an error term ε_i to model the difference between realized and expected values of the components of Y_i, we can write

$$Y_i = (\mu_A, \mu_T)\, X_i + \varepsilon_i$$

where $\varepsilon_i \sim (0, X_i' \Sigma X_i)$ and $\mathrm{var}(Y_i) = X_i' \Sigma X_i$.

Following Woodruff's (1989) paper, we summarize all available relevant information (population data, sample data, and the stochastic relationships between these data items) for all n units in the sample with data for at least one target variable in a linear model:

$$[Y_1,\ldots, Y_n] = [\mu_A, \mu_T][X_1,\ldots,X_n] + [\varepsilon_1,\ldots,\varepsilon_n]$$

$$\text{where } [\varepsilon_1,\ldots,\varepsilon_n] \sim \left[0 , \sum\nolimits_Y \right] \qquad (1)$$

or, writing this in a more compact form:

$Y = \mu X + \varepsilon$ where

ε has mean zero and covariance matrix Σ_Y and

Σ_Y is the diagonal matrix of the $\{X_i' \Sigma X_i\}$, where

Σ is the covariance matrix of (A_i, T_i).

Any known data relationships can be included in (1) through the covariance matrix Σ. For the Bureau's CES survey, a well–documented superpopulation model places strong restrictions on the form of Σ.

The linear relation given by (1) is the summary model. It would be possible to compute the generalized least squares (GLS) estimator of μ from this expression and extract the $\hat{\mu}_T$–components to estimate μ_T. It is often the case, however, that the target variables T_i have a conditional relationship, $E(T_i|A_i)$, that can be used to improve estimation. Auxiliary variables, data items known for every unit in the population, are used to adjust for the particular sample selected and the responding items in this sample. For noninformative sampling designs (Cassel, Särndal, and Wretman, 1977) the sample indicator variables, conditional on the sample outcomes of the auxiliary variables, are independent of the target variables. Hence, conditional on the auxiliary variable outcomes, the sampling distribution is irrelevant to estimation of the target variable means for such designs. In the case of CES, the sampling distribution is unknown because a non–probability design is used. It is fortunate and necessary that we can condition on an auxiliary variable and thereby compensate for most of the influence that the sampling mechanism may have on estimation.

We adjust for the sample selected by regressing on the auxiliary variable, exactly analogous to the univariate regression estimator (Cochran, 1977). The least squares linear regression estimator of T, a single target variable, on A, a single auxiliary variable, is

$$\hat{T} = \overline{T} + (\sigma_{TA} / \sigma_A^2)\, [\mu_A - \overline{A}]$$

In the multivariate case, we partition the covariance matrix of (A_i, T_i) to reflect the covariances between the auxiliaries (to which α refers) and targets (to which β refers) comparable to σ_{TA} and σ_A^2 above, writing it as

$$\Sigma = \begin{bmatrix} \Sigma_\alpha & \Sigma_{\alpha\beta} \\ \Sigma_{\beta\alpha} & \Sigma_\beta \end{bmatrix}.$$

Just as the univariate regression estimator conditions on the known auxiliary data, we can transform the target variables as

$$Z_i = [T_i - (A_i - \mu_A)\, \Sigma_\alpha^{-1}\, \Sigma_{\alpha\beta}]\, \chi_i.$$

The conditional mean and variance of Z_i are $E(Z_i \mid A_i) = \mu_T\, \chi_i$ and $V(Z_i \mid A_i) = \chi_i' (\Sigma_\beta - \Sigma_{\beta\alpha}\, \Sigma_\alpha^{-1}\, \Sigma_{\alpha\beta})\, \chi_i = \chi_i' \Sigma_\delta \chi_i$, where the last equality defines Σ_δ. We then let δ_i model the difference between realized and expected values

of the components of $(Z_i \mid A_i)$, writing

$$Z_i = \mu_T \chi_i + \delta_i.$$

Summarizing this over all sample units, we have

$$[Z_1,..., Z_n] = \mu_T [\chi_1,..., \chi_n] + [\delta_1,..., \delta_n]$$

$$\text{where } [\delta_1,..., \delta_n] \sim \left[0 , \sum\nolimits_Z \right] \qquad (2)$$

or, in a more compact form,

$$Z = \mu_T \chi + \delta \qquad \text{where}$$

$\chi = [\chi_1,..., \chi_n]$,

δ has mean zero and covariance matrix Σ_z,

Σ_z is the diagonal matrix of the $\{\chi_i \Sigma_\delta \chi_i'\}$,

and

$$\Sigma_\delta = \Sigma_\beta - \Sigma_{\beta\alpha} \Sigma_\alpha^{-1} \Sigma_{\alpha\beta},$$

From (2), a GLS estimator for μ_T, together with its variance, is

$$\hat{\mu}_T = Z \Sigma_z^{-1} \chi' (\chi \Sigma_z^{-1} \chi')^{-1}$$

$$\text{with variance } (\chi \Sigma_z^{-1} \chi')^{-1}$$

Although we don't know Σ_z, it is a function of Σ, the covariance matrix of (A_i, T_i). Σ can be accurately estimated by using the Markov superpopulation model, to be defined below, that describes CES data.

For this CES application, one auxiliary variable and four target variables were used. The four target variables for an establishment in month j are its employment in the previous four months, months j–4, j–3, j–2, and j–1. The auxiliary variable is the establishment's employment in month j–5. Although it is not technically an auxiliary variable, special features of CES data flow, explained in the last paragraphs of this section, allow us to treat employment in month j–5 as such.

Next we need Σ, the covariance matrix of (A_i, T_i). The model relating the components of $(A_i, T_i) = (A_i, T_{i1}, T_{i2}, T_{i3}, T_{i4})$ follows:

For all i, let

$$A_i = \beta_0 + \lambda_{i0},$$

$$T_{i1} = \beta_1 A_i + \lambda_{i1}, \text{ and}$$

$$T_{ij} = \beta_j T_{i,j-1} + \lambda_{ij} \text{ for } 2 \le j \le 4.$$

The $\{\beta_j\}$ are unknown constants. The $\{(\lambda_{i0}, \lambda_{i1},.., \lambda_{i4}) : 1 \le i \le N\}$ are iid random vectors with mean zero and diagonal covariance matrix. The variance of λ_{ij} exists for all j and thus

$$\tau_j = E(V(T_{ij}|T_{i,j-1})) = E(V(\lambda_{ij}|T_{i,j-1}))$$

exists for all i and $j \ge 1$. Let $\tau_0 = \sigma_0^2$ be the variance of A_i.

It follows from the above superpopulation model that if we denote the diagonal entries of Σ as $\{\sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\}$, where $\sigma_j^2 = V(T_{ij})$ for all i and for $1 \le j \le 4$, then off–diagonal entries are given by

$$(\Sigma)_{ik} = \sigma_{i-1}^2 \prod_{j=1}^{k-1} \beta_j \quad \text{for } 1 \le i \le 5,\ i < k \le 5,$$

where the $\{\beta_i\}$ are given above.

The inverse of this covariance matrix is tridiagonal with diagonal entries given by

$$(\Sigma^{-1})_{jj} = 1 / \tau_{j-1} + \beta_j^2 / \tau_j \text{ for } 1 \le j \le 4 \qquad (*)$$

except for the last diagonal entry, which is

$$(\Sigma^{-1})_{4,4} = 1 / \tau_4 , \qquad (**)$$

and immediate off–diagonal entries given by

$$(\Sigma^{-1})_{j,j+1} = (\Sigma^{-1})_{j+1,j} = -\beta_j / \tau_j$$

$$\text{for } 1 \le j \le 4 \qquad (***)$$

All other entries are zero.

The coefficients $\{\beta_j\}$ relate adjacent months' employment data and are estimated with the ratios of employment totals for those two adjacent months from sample units reporting for both months. Recall from section 2 that the links $\{\hat{\beta}_i\}$ are used to estimate these $\{\beta_j\}$. The variance of A_i, τ_0, is estimated from the benchmark historical data, and for $j \ge 0$, τ_j is estimated from the $\{\hat{\beta}_j\}$ by noting that for all i, $\tau_j = E[E((T_{ij} - \beta_j T_{i,j-1})^2 | T_{i,j-1})]$ is a quadratic in β_j. In particular, we have observed from historical CES data that τ_j is approximately at a minimum when $\beta_j = 1$ and increases as β_j moves away from 1 in either direction. For the population being considered in the simulation study of section 4, $\tau_0 \approx 10{,}000$ and $\tau_j \approx 28{,}000\ \beta_j^2 - 56{,}000\ \beta_j + 28{,}060$. Substituting the estimated links $\{\hat{\beta}_j\}$ into this expression, we obtain the estimates $\{\hat{\tau}_j\}$. The starred equations above yield an estimate of Σ^{-1}, and inverting it produces $\hat{\Sigma}$.

In our CES application, μ_A is unknown, as is the value of A_i for some sample units. Fortunately, good estimates of both are available for the case of the CES survey, as we next explain. Supposing that the current month is March, the vector (A_i, T_i) for establishment i consists of its employment values for October through February. A_i is the employment for October, while T_i is made up of the employment for months November through February. By March, 95% of the sample units will have reported their November employment, with similarly high cumulative response rates attained for the months preceding November. Given this relatively high response for months prior to November, the links for these long–past months will be based on about 90% or better of the sample units, and the LR estimator will approximately reduce to the simple ratio estimator for November and prior months. That is, it will do an excellent job at estimating total employment for five months or more into the past, and in particular it provides a good estimate of μ_A. This property of the LR estimator under CES data flow will allow us to use the estimated employment for the fifth month back as a pseudo–auxiliary variable.

Because of nonresponse, the auxiliary variable may not be available for each sample unit. We use a backwards predictor from the target variables that are observed for the particular sample unit to impute for a missing auxiliary variable. This predictor uses the reciprocals of the same links used in the LR estimator, linking backward from the responding target closest (in time) to the auxiliary variable.

For example, consider a sample unit with data missing for one and four months ago and available for two and three months ago, for which the response vector is (0, 1, 1, 0).

Continuing with the set–up of the previous example, this sample unit has data for December and January but no data for November and February. Suppose this sample unit has no data for the auxiliary variable, October employment. We use the estimates of the October–November and November–December links (say, $\hat{\beta}_{ON}$ and $\hat{\beta}_{ND}$) to predict the unit's auxiliary variable with $1 / \hat{\beta}_{ON} \times 1 / \hat{\beta}_{ND} \times T_D$ where T_D is the December employment for the unit.

Thus, establishment employment from five or more months in the past can be treated as an auxiliary variable because we can predict the missing employment data, and the LR estimator can be used to estimate the auxiliary variable population mean. This completes the definition of the L90 estimator with four target variables. A L90 with m targets is defined in an analogous manner.

4. A Simulation Study

A file containing six years of CES microdata from some 300,000 business establishments is available for research purposes. From this file we extracted 27 months (March 1986 to May 1988) of employment data on 3108 department stores for our initial simulation study. We extracted data only for those units that had reported their employment for all 27 months. To the extent that such conscientious reporting is anomalous, these units do not represent the actual CES universe. We then stratified them into two groups, those reporting fewer than 500 employees in March 1986 (2911 units) and those reporting more than 500 (197 units).

Twenty–seven months is the length of a complete CES estimation cycle. A new estimation cycle begins each June, at which time a newer set of benchmark employment counts becomes available. Estimates which had been computed by linking forward from the March two years and three months in the past are recomputed by linking forward from the newer set of benchmark counts, which refer to the March one year and three months ago. From June until the following May, the estimates link forward from these benchmark counts. The cycle re–starts the following June with the next set of newly available March benchmark counts.

The simulation results are tabulated for June 1987 to May 1988. During this time interval, the LR estimator used March 1986 benchmark counts. Prior to the arrival of the 1986 benchmark data, LR estimates are computed by linking forward from March 1985. After the March 1986 benchmark data are available, the employment estimates for April 1986 through March 1987 are re–calculated by linking forward from March 1986 data, using links computed from more nearly complete sample data (typically more than 90% complete). These re–benchmarked estimates are called fourth closing estimates. Based on nearly complete data, these fourth closing LR estimates for April 1986 to March 1987 are quite good, and the L90 does not attempt to compete with them. Only the first, second, and third closing estimates for June 1987 to May 1988, which must use less complete data, are of interest here, and it is here that the L90 and LR are tabulated.

Data arrive in a piecemeal fashion, making it necessary to update the three most current estimates of total employment each month. Recall that for each sample unit and month we have a closing code between one and five. A closing code of one (first closing) means that the sample unit's employment report for the given month arrived at the Bureau by the 12^{th} of the following month (that is, by the first closing date). A code of two means that the report arrived within the month after the first closing date (that is, by the second closing date). A code of three means that the report arrived during the next month (two months late), four that it came in after that but before the following March, and five that it came in some time after the following March (for all practical purposes, a non–respondent). Our use of closing codes of four and five departs from actual CES practice: in the simulation, a code of four indicates a unit for which data arrived three months late, while a code of five indicates one for which data never arrived. For example, first closing employment estimates for August are computed soon after the first of September and incorporate first closing data for August, first and second closing data for July, first, second, and third closing data for June, and all data not coded five for months prior to June. In September, first closing estimates for August are computed, second closing estimates for July, and third closing estimates for June. These third closing estimates are based on nearly complete data.

The simulation studies attempt to capture the data flow and estimation process for June 1987 to May 1988. To simulate the process, we used the sample file of 2911 department stores with March 1986 employment below 500 as a universe and selected a systematic sample of 485 units after sorting the file by March 1986 employment. Recall that the CES sample remains substantially fixed over time, aside from attrition and replenishment, and does not result from any known probability design. This sample was fixed for all replications of the estimation process, and simulation variability was generated by random generation of closing codes for all 27 months and for all 485 sample establishments.

An important problem for CES estimation is delayed reporting, especially on the part of large units. This means that first closing estimates are often based on data from only half the sample units. Second closing estimates are usually better because they are usually based on data from 70% or more of the sample units. The L90 was developed to improve first and second closing estimates.

We used the fixed sample of 485 units in replicating the data flow and estimation cycle from March 1986 to May 1988, generating for each replication fresh closing codes for each unit for each of the 27 months. The closing code probabilities for three response mechanisms (RM 1 – RM 3) are given in Table 1. In the CES survey, small units tend to respond before larger units. The random mechanisms used to generate closing codes attempt to mimic this phenomenon by giving different closing code distributions to sample units depending on their March 1986 employment (whether ≥ 200 or < 200).

For each replicate of a full set of closing codes on the 485 sample units, we calculate first, second, and third closing estimates for both the LR and L90 for each month from June 1987 to May 1988. Note that these estimates depend on fourth closing estimates computed for April 1986 through May 1987, which are also computed for each replicate. We can estimate mean square error (MSE) by comparing these estimates with the actual population values from all 2911 units.

For each response mechanism, we summarize these results in a table. Tables 2, 3, and 4 give ratios by month and closing of estimated MSE(L90) to estimated MSE(LR). The second half of each of these tables gives, for each month, the ratio of the revision for the L90 to that of the LR as well as the proportion of times this revision was smaller for the L90 than for the LR. These statistics are explained in greater detail below.

The first response mechanism, RM 1, is an extreme case. In practice, units seldom respond to the CES survey this slowly. RM 2 is a good approximation to usual response behavior. RM 3 is another extreme case, where the difference in response behavior between small and large firms is more severe than we have observed historically.

Table 2 summarizes the results of 50 replications of CES data flow and estimation under RM 1 (fourth closing estimates for March 1986 to May 1987; first, second, and third closing estimates for June 1987 to May 1988). The

ratios $\hat{MSE}(L90) / \hat{MSE}(LR)$ are tabulated for every other month from June 1987 to May 1988, and for first, second, and third closing (MSE behavior for months not tabulated is essentially identical and is omitted to avoid data clutter). This table shows the greatest gain from using the L90. This is just as the theory predicts; the L90 thrives on nonresponse, compared to the LR.

Tables 3 and 4 are based on 100 replications of RM 2 and RM 3 respectively. The MSE ratios are closer to one in these tables, where far less nonresponse is modeled. For all three response mechanisms, the number of replications was sufficient to assure that a 99% confidence interval for MSE(L90) / MSE(LR) did not include the value 1. Note that by third closing the MSE of the two estimators is similar.

The second part of each of tables 2, 3, and 4 summarizes ratios of average first–to–second closing absolute revisions and relative incidence of small revisions for each month. For example, under RM 2 for September 1987, Table 3 shows that 60 of the 100 absolute first–to–second closing revisions for September 1987 employment estimates were smaller for the L90 than for the LR. Table 3 also shows that for September 1987 the average absolute revision for the L90 was 3/4 of the average absolute revision for the LR.

Table 5 summarizes the incidence of large first–to–second closing revisions for each response mechanism and estimator. A revision is defined as large if it is greater than 1% of total employment. For example, for RM 2 we have 1100 first–to–second closing revisions (100 replications by 11 revisions — no second closing estimates for months after May 1988 were computed, so there were first–to–second revisions for June 1987 to April 1988 only). Of the 1100 revisions, in only one case were both LR and L90 large. In 33 cases L90 was small and LR large; in six, the other way around. Both produced small revisions in the remaining 1060 cases.

These incidences of large revisions show the most extreme differences between the two estimators. In summary, for every large L90 revision one should expect five or six large LR revisions.

Table 1. Closing Code Probabilities

Closing Code		1	2	3	4	5
Response Mechanism	Employment Category					
1	< 200	.30	.25	.30	.10	.05
	≥ 200	.25	.25	.25	.20	.05
2	< 200	.55	.15	.20	.05	.05
	≥ 200	.45	.20	.25	.05	.05
3	< 200	.60	.15	.10	.06	.09
	≥ 200	.35	.25	.25	.10	.05

Table 2. Estimated Mean Square Error Ratios and Revision Ratios and Rates by Month Response Mechanism 1

	Jul 87	Sep 87	Nov 87	Jan 88	Mar 88	May 88
MSE by closing						
first	.60	.68	.51	.38	.68	.48
second	.95	.98	.86	.97	.76	.66
third	.97	1.04	1.01	1.01	.86	.91
Revision						
ratios	.67	.64	.62	.56	.80	
rates	.64	.70	.72	.74	.62	

Table 3. Estimated Mean Square Error Ratios and Revision Ratios and Rates by Month Response Mechanism 2

	Jul 87	Sep 87	Nov 87	Jan 88	Mar 88	May 88
MSE by closing						
first	.67	.90	.79	.65	.87	.88
second	.88	.89	.91	.97	.85	.92
third	.98	.97	.99	.99	.93	1.03
Revision						
ratios	.62	.60	.57	.66	.62	
rates	.74	.75	.76	.71	.78	

Table 4. Estimated Mean Square Error Ratios and Revision Ratios and Rates by Month Response Mechanism 3

	Jul 87	Sep 87	Nov 87	Jan 88	Mar 88	May 88
MSE by closing						
first	.74	.93	.75	.69	.75	.83
second	.98	.93	.94	.89	.67	.87
third	.99	.97	.98	.96	.86	1.0
Revision						
ratios	.53	.60	.58	.69	.57	
rates	.87	.80	.72	.69	.84	

Table 5. Incidence of Large Revisions by Estimator

	LR					
	Response Mechanism 1		Response Mechanism 2		Response Mechanism 3	
	Large	Small	Large	Small	Large	Small
L90						
Large	30	15	1	6	2	7
Small	90	415	33	1060	44	1047
Total	120	430	34	1066	46	1054

Totals Across Mechanisms

	LR		
L90	Large	Small	Total
Large	33	28	61
Small	167	2522	2689
Total	200	2550	2750

5. Conclusions

This paper develops a concrete application of a multivariate estimation technique for minimizing the negative effects of nonresponse in sample surveys. The technique makes use of the multivariate relationships between different data items to develop an estimator, the L90, for the Bureau's Current Employment Statistics survey.

This work extends some theoretical estimation work, Woodruff (1989), to estimation under the far greater complexities of CES data flow.

These complexities introduce parameter estimation problems for which workable solutions were found (section 3). These solutions may themselves be of theoretical interest, as they extend slightly the definition of auxiliary variable (and covariate), or rather muddy the difference between target and auxiliary variables.

If the bottom line is the reduction of mean square error, then the L90 is an improved CES estimator. It makes better use of both the available data and the known relationships between data items and results in a reduction in variance and closing revisions. Unlike the LR, it also provides us with a measure of its variability.

The current CES estimator, the LR, has been in use since the 1940's. It is simple and intuitive, and it does an excellent job, despite minor shortcomings. It is, in fact, a special case of the L90 estimator with T_i consisting of a single target variable. Its shortcomings can yield to the revolution in computing power that has occurred in the decades since the LR was developed. The L90 estimator would have been a computational impracticality at the time the LR was adopted by the CES survey.

The testing described in the simulation section of this paper is still preliminary. Results are encouraging, but similar results need to be obtained for many industries and size classes. Other factors not considered here were the effects of sample imbalance and other response mechanisms. Finally, the LR and the L90 need to be run in parallel and compared in monthly production runs of CES estimates.

The L90 provides greater flexibility than the LR for coping with several recurrent CES estimation problems. Some of the problems are these:

1) Response probabilities strongly correlated to the quantities being measured. The effect of these correlations on bias in the estimates can be reduced by adapting the L90 along the lines given in Woodruff (1989).

2) The need for estimates of precision of published CES estimates. Because the L90 is a GLS estimator, GLS methodology gives an estimate of its variance.

3) Adaptability of the estimator to varying rates of data flow. It is easy to vary the number of target variables in the L90 by estimation cell to achieve an optimal fit. This may prove efficient as improvements in data collection increase first closing response rates.

4) Ability of the estimator to measure sudden economic shifts. The L90 is better at picking up such changes than the LR.

Some considerations for further research include these:

1) Bias adjustment for the LR and the L90. Both estimators appear to be unbiased under a fixed universe. Most of the historically observed bias in these estimators probably comes from births and deaths in the population of business establishments.

2) The effect of introducing probability sampling techniques on the series of CES estimates.

3) The magnitude of revisions of the L90 in cases where those of the LR are large.

BIBLIOGRAPHY

Cochran, W. (1977), *Sampling Techniques*, New York: Wiley & Sons.

Grzesiak, T., and Copeland, K. (1987), "Evaluation of Estimators for Employment Level and Change in Establishment Surveys," ASA Proceedings of the Section on Survey Research Methods.

Madow, L., and Madow, W. (1978), "On Link Relative Estimators," ASA Proceedings of the Section on Survey Research Methods, 534–539.

Pfeffermann, D. (1988), "The Effect of Sampling Design and Response Mechanism on Multivariate Regression–based Predictors," *Journal of the American Statistical Association*, **83**, 824–833.

Royall, R. (1981), "Study of Role of Probability Models in 790 Survey Design and Estimation," BLS technical report.

Royall, R., and Cumberland, D. (1981), "An Empirical Study of the Ratio Estimator and Estimators of Variance," *Journal of the American Statistical Association*, **76**, 66–77.

Rubin, D., and Little, R. (1988), *Statistical Analysis with Missing Data*, New York: Wiley & Sons.

West, S. (1982), "Linear Models for All Employment Data," BLS report.

West, S. (1983), "A Comparison of Different Ratio and Regression Type Estimators for the Total of a Finite Population," ASA Proceedings of the Section on Survey Research Methods.

Woodruff, S. (1982), "A Comparison of Some Estimators in Sampling for a Time Series with Linear Trend, " ASA Proceedings of the Section on Survey Research Methods.

Woodruff, S. (1983), "Variance Estimation for a Time Series with Linear Trend, " ASA Proceedings of the Section on Survey Research Methods.

Woodruff, S. (1989), "Estimation in the Presence of Nonignorable Missing Data and a Markov Superpopulation Model, " ASA Proceedings of the Section on Survey Research Methods.