ABSTRACT

Title of Dissertation: **PANEL SURVEY ESTIMATION IN THE PRESENCE OF LATE REPORTING AND NONRESPONSE**

Kennon R. Copeland, Doctor of Philosophy, 2004

Dissertation Directed By: **Professor Partha Lahiri**
Joint Program in Survey Methodology

Estimates from economic panel surveys are generally required to be published soon after the survey reference period, resulting in missing data due to late reporting as well as nonresponse. Estimators currently in use make some attempt to correct for the impact of missing data. However, these approaches tend to simplify the assumed nature of the missing data and often ignore a portion of the reported data for the reference period. Discrepancies between preliminary and revised estimates highlight the inability of the estimation methodology to correct for all error due to late reporting.

The current model for one economic panel survey, the Current Employment Statistics survey, is examined to identify factors related to potential model misspecification error, leading to identification of an extended model. An approach is developed to utilize all reported data from the current and prior reference periods, through missing data imputation. Two alternatives to the current models that assume growth rates are related to recent reported data and reporting patterns are developed, one a simple proportional model, the other a hierarchical fixed effects model.

Estimation under the models is carried out and performance compared to that of the current estimator through use of historical data from the survey. Results, although not statistically significant, suggest the potential associated with use of reported data from recent time periods in the working model, especially for smaller establishments.

A logistic model for predicting likelihood of late reporting for sample units that did not report for preliminary estimates is also developed. The model uses a combination of operational, respondent, and environmental factors identified from a reporting pattern profile. Predicted conditional late reporting rates obtained under the model are compared to actual rates through use of historical information for the survey. Results indicate the appropriateness of the parameters chosen and general ability of the model to predict final reporting status. Such a model has the potential to provide information to survey managers for addressing late reporting and nonresponse.

PANEL SURVEY ESTIMATION IN THE PRESENCE OF LATE REPORTING AND NONRESPONSE

By

Kennon R. Copeland

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2004

Advisory Committee:
Professor Partha Lahiri, Chair
Professor Katherine Abraham
Professor John Eltinge
Professor Nathaniel Schenker
Professor Eric Slud
Professor Richard Valliant

Dedication

This dissertation is dedicated to my children, Brian and Erin Copeland, and to the memory of my father, Doyal L. Copeland.

Acknowledgements

There are many individuals who played a significant role in my being able to complete this degree, and I wish to recognize some of them here.

Partha Lahiri, for his guidance and support in carrying out the research and writing the dissertation.

Bob Groves, for having faith in my capability and encouraging me to strive for this degree.

Roger Tourangeau, for providing ongoing support and encouragement from JPSM.

Paul Wilson, for his support in my initiating this journey and for allowing me the time to do the research.

Rick Valliant, for his input helping me to clarify my research problem.

John Eltinge for his guidance in exploring the problem of interest.

Cathy Dippo and Steve Cohen, for welcoming me back to BLS to carry out my research.

Katherine Abraham, Nat Schenker, and Eric Slud, for their suggestions along the way.

Nancy Mathiowetz, for her insights about getting through the process.

Trivellore Raghunathan, for his advice when I started down my research path.

Pam Ainsworth, Rupa Jethwa, Adam Kelley, and Sarah Dipko, for their support in working through all the administrative, technical, and programming difficulties.

Sunghee Lee and Scott Fricker, for their collegiality and support as fellow members of the first JPSM PhD cohort.

Table of Contents

List of Tables

List of Figures

Chapter I: Overview

A. Introduction

Many economic surveys must strike a balance between timeliness and accuracy in the generation of estimates. Estimates are generally required to be published soon after the survey reference period in order to efficiently guide policy aimed at affecting the marketplace. Speed of delivery can adversely affect survey quality, however, as nonreporting will tend to be higher with shorter collection periods. Estimation methods developed for these surveys are intended to compensate for missing data so as to reduce the error due to nonreporting.

A portion of survey nonreporting within such a survey environment can often be viewed as temporal, with responses for some sample units becoming available subsequent to the prescribed collection period (referred to here as “late reporting”). The remaining portion of survey nonreporting reflects sample units that never report data for the reference period (referred to here as “nonresponse”). One approach commonly taken with economic data series is the issuance of preliminary estimates shortly after the reference period, based upon sample data received within the prescribed collection period (referred to here as “preliminary reporting”), followed by one or more revised estimates based upon data from both preliminary and late reporters.

Despite the issuance of revised estimates, preliminary estimates are most critical for use and tend to receive the most visibility. Deviations between preliminary and revised estimates may be perceived as an inability of the estimation methodology to appropriately correct for nonreporting. Although information on sampling and other errors associated with the preliminary estimates may be provided, and may show revisions are not outside the bounds of expected survey error, the perception of survey performance may still be tied to the nature of differences between preliminary and revised estimates. This is especially true when looking at revisions to period-to-period change in the estimates, where a difference between preliminary and revised estimates deemed inconsequential for the reference period level may be greater than the estimate of period-to-period change. Thus one key objective for such surveys is reducing the potential for large differences between preliminary and revised estimates, both level and change.

B. Discussion of Problem

Estimators currently in use for economic panel surveys of establishments often utilize relationships between current and prior period values in deriving current period estimates, in part to control variability in period-to-period change that would result from differences in the set of reporting sample units from one period to the next. As a result, these estimators may restrict usable sample to those units reporting data for both periods. The set of usable sample units will expand between the generation of preliminary and revised estimates with the addition of late reporters for the current period which had reported for the prior period. For example both the Current Employment Statistics survey, conducted by the Bureau of Labor Statistics, and the Monthly Retail Trade survey, conducted by the Census Bureau, revise estimates based upon late reporters.

The magnitude of the difference between preliminary and revised estimates depends in part upon the extent to which preliminary reporters can be used as proxies to reflect the distribution for late reporters. To the extent distributions deviate for late reporters, revised estimates will show larger differences from the preliminary estimates. The potential for different distributions may be exacerbated when using estimators which utilize both current and prior period values in deriving current period estimates.

Estimators often attempt to control for the impact of missing data due to late reporting and nonresponse through the creation of estimation cells, defined primarily through the use of information available for the entire population, in which nonreporting is assumed to be random (i.e., that within an estimation cell preliminary reporters reflect the relationship between current and prior period values for late reporters).

There are two key issues associated with estimation methods currently used that bear consideration in developing methods intended to reduce differences between preliminary and revised estimates:

1) Discarding data from sample that fail to report for both the current and immediately prior reference period. A portion of the data discarded when generating preliminary estimates ends up being used when generating revised estimates, that being prior period data for current period late reporters. By developing approaches for preliminary estimation that make direct use of data that may be included in the revised estimates, differences between preliminary and revised estimates could potentially be reduced.

2) Assuming no difference in relationship between current and prior month regardless of prior reporting patterns. The underlying models used for current

estimation methods typically assume relationships that depend only upon information known for the population. Prior data on sample units, which will be more complete for consistent reporters than for sample units that report sporadically, may provide insight into the relationship between current and prior periods.

An approach to address these issues and potentially reduce the magnitude of differences between preliminary and revised estimates would be to expand the model underlying the current estimators to encompass differential relationships based upon prior reporting patterns and available data, and to impute for missing current period data when prior period data are present. That is the approach taken in this dissertation research.

C. Statement of Purpose

The primary objective of this dissertation research was to develop an estimation approach for panel surveys, given late reporting and nonresponse, yielding improved accuracy for preliminary estimates of monthly population totals and month-to-month change in population totals, relative to that achieved by estimators currently in practice. The focus was on more complete and effective use of available population and sample information than is currently the case, through imputation for missing data due to late reporting and nonresponse, so as to reduce the difference between preliminary and revised estimates. An example panel survey, the Current Employment Statistics (CES) survey, was used for developing alternative approaches and for assessment of performance.

The performance of estimates resulting from the working models were compared to that for estimates derived by the methodology currently in practice, by comparing differences between preliminary and revised estimates. The focus was on late reporting and nonresponse error. Measurement error, although important in addressing the overall accuracy of survey estimates, was not addressed in this research.

A secondary objective of this dissertation research was to develop a model for predicting final reporting status for sample units other than preliminary reporters. A logit model appears appropriate for use in this regard, with independent variables selected on the basis of late reporting and nonresponse patterns. The working model was developed so as to balance parsimony and incorporation of relevant factors and information.

D. Statement of Work

In the dissertation research, the following activities were carried out:

1. Review the statistical literature relative to late reporting and nonresponse, especially as it applies to panel surveys of establishments (Chapter II);
2. Describe the survey design and estimation methods currently used in an example panel survey subject to late reporting and nonresponse, so as to motivate the research problem (Chapter III);
3. Analyze the example panel survey in terms of reporting patterns and develop and assess a model for predicting final reporting status (late reporter, nonresponse) for sample establishments failing to report during the survey's preliminary reporting period (Chapter IV).

4. Analyze the example panel survey in terms of impact of late reporting and nonresponse and develop and assess working models for imputing missing employment values for sample establishments reporting only one of the current and prior months and for estimating current month employment (Chapter V);
5. Comment on the implications of the research findings as they relate to future research and implementation in a survey setting (Chapter VI).

Chapter II: Literature Review

A. Introduction

Unit nonresponse is a common occurrence in sample surveys that, if ignored, results in increased variance and likely bias for survey estimates. Nonresponse decreases the effective sample size of the survey, thereby increasing the variance of survey estimates. In addition, nonresponse can yield biased estimates if the distribution for variables of interest for respondents differs from that for nonrespondents. As discussed in Chapter I, an additional aspect of unit nonresponse in surveys (both cross-sectional and panel) is that related to late reporting (see e.g., Bureau of Labor Statistics, 2004a; Hogan, et al., 1997).

As noted by the Committee on National Statistics' Panel on Incomplete Data (Madow, et al., 1983), "…the inevitable nonresponse requires the consideration of methods that improve analysis by statistical adjustment of the collected data. But no statistical methods will fully compensate for missing units and data. Biases will almost certainly remain. Good methods are chiefly aimed at reducing biases and mean square error of estimators while reducing or at least not unduly increasing variances of estimators."

A common approach to compensating for unit nonresponse in cross-sectional surveys is through weighting adjustments utilizing auxiliary information about the sample units from the frame. A number of books, monographs, and papers addressing general methods and theory of weighting adjustments for unit nonresponse have been written (see, e.g., Little and Rubin, 2002; Oh and Scheuren, 1983; Kalton and Kasprzyk, 1986; and, in the establishment survey setting, Hidiroglou, et al., 1995).

Unit nonresponse within panel surveys takes on an additional dimension beyond that for cross-sectional surveys (i.e., time). Although some sample units will be nonrespondents for all survey periods and others will provide responses for all survey periods, many sample units will respond for some survey periods and be nonrespondents for others. Panel surveys thus offer the potential for a richer set of auxiliary information (i.e., values for the variables of interest for the nonresponding sample unit from prior and/or succeeding survey periods) on which to base a nonresponse compensation method than that available from an analogous cross-sectional survey, albeit at the price of more complex reporting patterns. This environment not only allows for a wider range of weighting adjustment methods to be considered for panel surveys, but also makes imputation a more desirable option.

In spite of the availability of additional auxiliary data and the long existence of panel surveys, compensation for nonresponse in panel surveys is often based on cross-sectional methods or some variant developed to fit panel surveys. In addition, much of the nonresponse literature is focused on cross-sectional surveys. For example, the recent nonresponse text by Groves and Couper (1998) does not address the issue of nonresponse in a longitudinal survey other than that occurring in the first wave.

This chapter presents a discussion of nonresponse adjustment methods as applicable to panel surveys. As a framework, the chapter begins with a brief overview of panel surveys, followed by a discussion of key nonresponse theory and common adjustment methods used for cross-sectional surveys. A mathematical framework for unit nonresponse in panel surveys is then presented, followed by a review of adjustment methods currently in use for panel surveys. The issue of late reporting in panel surveys is then discussed. The chapter closes with a discussion of future direction for this area of research.

B. Panel Surveys

Panel surveys are "surveys in which similar measurements are made on the same sample at different points in time" (Kasprzyk, et al., 1989). Panel surveys may involve complete overlap of the sample across time, rotation of the sample units across time, or a combination of complete overlap and rotation of the sample across time.

Duncan and Kalton (1987) discuss characteristics of surveys across time – panel and repeated surveys. Although both panel and repeated surveys provide estimates for a population at multiple points in time, panel surveys are particularly well-suited for estimating gross and other components of individual change and for aggregating data for individuals over time, important characteristics for use in economic analysis (Solon, 1989). In addition, panel surveys provide advantages for collecting data on events occurring in specified time periods and, with some mechanism for taking into account population changes, also allow for estimating net changes. Bailar (1989) and Binder (1998) provide discussions of key issues associated with surveys across time. One important issue is the types of estimates desired, e.g., if cross-sectional estimates are required in addition to estimates of change. Maintenance of an accurate sampling frame must be planned for. Respondent burden becomes critical, as sample units are expected to provide data multiple times. Sample attrition must be considered, not only because of the nonresponse effect for later time periods, but also because of the impact on selected analyses given missing time periods for a given sample unit. Finally, nonresponse adjustment is complicated as units may not respond for every time period.

Panel surveys differ from cross-sectional surveys in terms of the manner in which nonresponse may be classified and the information available about nonrespondents upon which to base approaches to compensate for nonresponse. As a result, approaches to compensating for nonresponse often differ for panel surveys from those for cross-sectional surveys.

C. Nonresponse Overview

Unit nonresponse is defined as "…a complete failure to obtain data from a sample unit…" (Office of Management and Budget, 2001). An obvious implication of unit nonresponse is a variance increase due to the reduction in the effective sample size. The variance increase for a sample mean from a simple random sample can be expressed (ignoring the finite population correction) as the ratio of the total to the responding sample size

$$\frac{Var(\bar{y}_r)}{Var(\bar{y}_n)} = \frac{S_y^2/(n-m)}{S_y^2/n} = \frac{n}{n-m}$$

where $\bar{y}_r =$ mean for sample respondents

$\bar{y}_n =$ mean for all sample units

$S_y^2 =$ population variance for Y

$n =$ total number of sample units

$m =$ number of nonrespondents

Based upon this relationship, one approach to alleviating the impact of nonresponse on survey variance is to oversample based upon the anticipated nonresponse rate, so the responding sample size is expected to be that required to achieve the target variance.

A second implication of unit nonresponse is the potential for biased estimates. Nonresponse bias for a sample mean from a simple random sample can be expressed as the product of two components: the nonresponse rate and the difference between the means of the respondents and the nonrespondents

$$Bias(\bar{y}_r) = E(\bar{y}_r - \bar{y}_n) = \frac{m}{n}(E(\bar{y}_r) - E(\bar{y}_m))$$

where $\bar{y}_m =$ mean for sample nonrespondents

$\frac{m}{n} =$ unit nonresponse rate

Reducing the unit nonresponse rate thus serves to lessen the bias implications of nonresponse as well as the variance, while reducing the difference between nonrespondents and respondents relative to the variables of interest should serve to lessen the nonresponse bias. Operational refinements regarding questionnaire design, collection mode, response burden, and survey protocol can be implemented in an attempt to reduce nonresponse rates (see e.g., Lessler and Kalsbeek, 1992). A combination of operational refinements, such as collection of observational information about nonresponding units, and statistical methods, such as weight adjustments or imputation, are typically applied to reduce the bias impact of nonresponse.

Development of statistical methods resulting in a decrease in the bias due to nonresponse requires an understanding of the response mechanism and the relationship between respondents and nonrespondents. The response mechanism is commonly viewed in terms of the reason for nonresponse: refusal; unavailable (referred to as "not-at-home" in the household survey setting); inability to participate; and not located (see, e.g., Kalton and Kasprzyk, 1986; Groves, 1989). Classification of nonrespondents in terms of reason for nonresponse requires information be collected as part of the survey protocol.

Several recent articles (Curtin, et al., 2000; Keeter, et al., 2000; Merkle and Edelman, 2002) have called into question traditional assumptions about the impact of higher nonresponse rates, or at the least the perceived benefits of reducing nonresponse rates. These articles suggest lowering nonresponse rates (i.e., reducing the first component of the nonresponse bias equation) may actually increase the second component of the nonresponse bias equation, the difference between the means of the respondents and the nonrespondents, leaving a net result of no gain in terms of nonresponse bias.

Panel surveys add another dimension to the response mechanism, that being response status at different points in time. Little and David (1983) distinguished three types of panel survey nonresponse – attrition (sample unit stops reporting), late entry (sample unit does not report initially), and reentry (sample unit has a gap in reporting). While this categorization describes general patterns of nonresponse, the types are not mutually exclusive. A sample unit that stops reporting (attrition) could have had gaps in reporting for time periods prior to attrition (reentry) and may not have reported initially (late entry).

Little and Su (1989) identified two patterns of panel survey nonresponse – monotone (the only type of nonresponse is attrition) and haphazard (nonresponse is either late entry or reentry or both). For a monotone pattern of nonresponse, if a sample unit is a nonrespondent for time period, t, then the sample unit is a nonrespondent for any time period $t^* > t$. Thus, under a monotone pattern of nonresponse the set of responding sample units for time period $t+1$ is a subset of the set of responding sample units for time period t. Although a fully monotone pattern of nonresponse is unlikely, the actual pattern may be approximately monotone (e.g., dropouts in clinical trials).

These two taxonomies could be refined to reflect more completely the nature of reporting patterns. The Little and David taxonomy ignores mixtures of patterns, while the haphazard response category of Little and Su's taxonomy does not provide useful distinctions among haphazard patterns, encompassing a wide variety of nonresponse patterns (e.g., the late entry and reentry panel nonresponse types described by Little and David, as well as any combination of Little and David's three nonresponse types). Clarifying the distinctions among patterns of nonresponse could prove useful in developing a nonresponse compensation method, as distributional properties may differ among patterns. In addition, complete nonresponse (sample unit never reports) should be added to the list of nonresponse types and complete response (sample unit always reports) should be added so all sample units are encompassed by the classification. As a result, it may be more appropriate to talk in terms of panel survey reporting patterns rather than nonresponse patterns. An expanded and refined taxonomy of reporting patterns for panel surveys is proposed in Chapter IV.

D. Nonresponse Models

Statistical methods to compensate for unit nonresponse require models, whether explicitly stated or implicitly assumed, specifying the relationship between respondents and nonrespondents in terms of available reported and auxiliary information. Although methods for compensating for unit nonresponse have been a part of survey methodology for over 50 years (see, e.g., Cochran, 1953; Hansen, et al., 1953), the last quarter century has seen the development of more rigorous theoretical foundations upon which to build survey-specific models for compensating for unit nonresponse. These foundations have provided an approach for explicitly stating underlying assumptions that often were implicitly assumed historically, and for selecting an appropriate method to compensate for nonresponse.

1. Ignorability

An important concept in determining the appropriateness of a model is ignorability, formulated by Rubin (1976). Ignorability may be viewed as defining the conditions under which the missing data mechanism does not depend upon missing values resulting from nonrespondents, and therefore inferences about the population can appropriately be made using only observed values i.e., the nonresponse mechanism and the missing data for nonrespondents can be “ignored” when making inferences (although ancillary information about the nonrespondents may be required).

The concept of ignorability can be stated in terms of conditional probability distributions (Little and Rubin, 2002). If Y represents the data for the variable of interest, which are subject to nonresponse, X represents ancillary data, which are fully observed, and M represents missingness of the data, then the joint distribution of $\left(X,Y\right)$ and M can be written as

$$f\left(X,Y,M\mid\theta,\phi\right)=f\left(X,Y\mid\theta\right)f\left(M\mid X,Y,\phi\right)$$

The missing data mechanism is characterized by the conditional distribution of M given $\left(X,Y\right)$, $f(M\mid X,Y,\phi)$, where ϕ denotes unknown parameters of the distribution.

Note that Y can be decomposed into observed, Y_{obs}, and missing, Y_{mis}, components, i.e., $Y=\left(Y_{obs},Y_{mis}\right)$.

If missingness does not depend on the values of the variable of interest regardless of status, i.e., if

$$f\left(M\mid X,Y,\phi\right)=f\left(M\mid\phi\right) \text{ for all } Y \text{ and } \phi$$

then the data are said to be missing completely at random (MCAR).

If missingness depends only on the components of Y that are observed, i.e., if

$$f\left(M\mid X,Y,\phi\right)=f\left(M\mid X,Y_{obs},\phi\right) \text{ for all } Y_{mis} \text{ and } \phi$$

then the data are said to be missing at random (MAR).

If missingness depends on components of Y that are missing, i.e., if

$$f\left(M\mid X,Y,\phi\right)=f\left(M\mid X,Y_{obs},Y_{mis},\phi\right) \text{ or}$$

$$f\left(M \mid X, Y, \phi\right) = f\left(M \mid X, Y_{mis}, \phi\right)$$

then the data are said to be not missing at random (NMAR).

In addition, the missing data mechanism is ignorable for data meeting either the MCAR or MAR condition, and for which the parameters of the joint distribution of $\left(X, Y\right)$ and M, θ and ϕ, are distinct, in the sense that the joint parameter space of $\left(\theta, \phi\right)$ is the product of the parameter space of θ and the parameter space of ϕ.

If the missing data mechanism is ignorable, the distribution of the missing values conditional on the observed values and the response mechanism is equivalent to the distribution of the missing values conditioned solely on the observed data and, therefore, unbiased estimates for nonrespondents may be derived based upon observed values along with ancillary information known for the sample. In practice, establishing a condition of MAR or MCAR will require the population may be segmented into groups such that the MAR condition is met (or approximately met) within each group.

2. Selection Models

Selection models fit closely with Rubin’s concept of ignorability. The joint distribution of Y and M given unknown distribution parameters θ and ψ can be factored as

$$f\left(Y, M \mid \theta, \psi\right) = f\left(Y \mid \theta\right) f\left(M \mid Y, \theta, \psi\right)$$

Developing a nonresponse adjustment approach can then be viewed as defining appropriate conditions under which the data can be viewed as MAR, thereby allowing inference from the observed data. Little (1986) discusses two common approaches to defining conditions for estimation of means—response propensity and predicted means. In both approaches, the objective is to define strata within which the data can be viewed as MAR. This is accomplished by stratifying the sample on some auxiliary variable, X, known for the population, for which the variable of interest, Y, is (believed to be) conditionally independent of the response status, $r\ (=0,\ 1)$.

The response propensity approach, suggested by David, et al. (1983), utilizes the propensity score theory of Rosenbaum and Rubin (1983). The response propensity given an auxiliary variable, X, is given by $p\left(X\right)=P\left(r=1|\,X\right)$. Under response propensity theory, if the auxiliary variable can be shown to be conditionally independent of the response indicator, r, given $p(X)$, then the variable of interest is also conditionally independent of the response indicator given $p(X)$. The result is the definition of conditions under which MAR holds and inference for the full population may be made from the observed data.

In practice, estimates of $p(X)$ are generated from the logistic regression of r on X, and nonresponse adjustment strata are formed based upon grouped values of the estimated $p(X)$, under the assumption that conditional independence holds within grouped values,. An example of an application of the response propensity approach for defining nonresponse adjustment cells is provided in Rizzo, et al. (1996).

Under the predicted means approach, the objective is to stratify the sample such that the distribution of Y is the same for respondents and nonrespondents within stratum. As the values of Y are not known for the entire sample, an auxiliary variable, X, correlated with Y and known for the sample, is used. In practice, estimates of $\bar{Y}$ are generated from the regression of Y on X, and nonresponse adjustment strata are then formed based upon grouped values of the estimated $\bar{Y}$, under the assumption respondents and nonrespondents share the same distribution within grouped values.

As discussed in Little (1986), response propensity stratification reduces large-sample bias (that portion of the bias which dominates the overall bias as the sample size increases), while predicted means stratification reduces both bias and variance. One drawback to predicted means stratification is that it requires separate models and nonresponse adjustments for each variable of interest to achieve the gains.

Both response propensity and predicted means approaches to defining nonresponse adjustment cells rely on the correlation between X and Y for inferring ignorability and upon the assumption that small deviations in distributions among units classified in the same cell do not adversely affect the assumption of ignorability. For surveys with large numbers of variables of interest, establishing nonresponse adjustment cells on the basis of a single (albeit possibly multivariate) X can strain the assumption that ignorability holds for each variable of interest. Surveys in which either the response propensity or the predicted means are continuous in nature are also subject to lack of robustness of the ignorability assumption.

3. Pattern-Mixture Models

Little (1993) proposed the use of pattern-mixture models for handling incomplete multivariate data, such as that arising from a panel survey. This approach differs from the selection model approach in the decomposition of the joint distribution of the observation matrix, Y, and missing-data indicator matrix, M. Pattern-mixture models invert the assumption concerning conditionality between Y and M, specifying that the distribution of Y is conditioned on the missing data pattern, M:

$$f(Y, M \mid \varphi, \pi) = f(M \mid \varphi) f(Y \mid M, \pi)$$

Separate models are then required for Y conditioned upon each missing data pattern. When the data are MCAR, the pattern-mixture model is equivalent to the selection model.

Pattern-mixture models lead to marginal distributions for Y that are mixtures of distributions (e.g., mixture of normal distributions, with different parameters for each missing data pattern, rather than one normal distribution with a consistent set of parameters across missing data patterns). These models are typically underidentified due to the missing data, requiring restrictions be specified to allow identification of all model parameters. In this sense, the pattern-mixture model approach can be viewed as a means of recognizing and addressing nonignorability of the response mechanism. Pattern-mixture models provide an approach for explicitly stating the assumptions about data relationships without the need for the fully restrictive assumptions of data assumed MAR.

Using Little's (1986) illustration, assume a survey is taken at two time periods, $t\ (=1,\ 2)$. There are four potential response patterns, $(r_1, \mathrm{r}_2) = (1,1), (1,0), (0,1), (0,0)$. Rather than assume the joint distribution of Y_1 and Y_2 is the same for all missing data patterns as with the complete data pattern, the pattern-mixture model approach allows specification of separate models for each missing data pattern. As can be seen, the conditional distributions $f\left(Y_2 \mid Y_1, (0,1)\right)$, and $f\left(Y_1 \mid Y_2, (1,0)\right)$, and the joint distribution $f\left(Y_2, Y_1 \mid (0,0)\right)$ cannot be estimated given the data.

Use of pattern-mixture models requires specification of models for each missing data pattern, as well as specification of models specifying the distribution of unidentified parameters (called "identifying restrictions"). Complete-case missing-variable (CCMV) restrictions equate all missing variables to the complete case pattern.

Returning to the illustration, CCMV restrictions would specify

$$f\left(Y_2 \mid Y_1, (0,1)\right) = f\left(Y_2 \mid Y_1, (1,1)\right)$$

$$f\left(Y_1 \mid Y_2, (1,0)\right) = f\left(Y_1 \mid Y_2, (1,1)\right)$$

$$f\left(Y_2, Y_1 \mid (0,0)\right) = f\left(Y_2, Y_1 \mid (1,1)\right)$$

This is analogous to the approach taken with selection models and, if all parameters (identifiable or not) for the models corresponding to missing data patterns are assumed equivalent to those for the complete case, will simplify to the MAR assumptions. The difference between the pattern-mixture model and the selection model under the CCMV restrictions is that parameters for missing data patterns can differ from those for complete cases in situations where the parameters are estimable. For panel surveys, this means prior information about the sample unit could be used to estimate the parameters of the assumed distribution, rather than having to rely solely on respondents from the current reference period. Other restrictions can be defined to fit expected relationships between missing data and estimable parameters.

For example, in panel surveys missing data patterns reflecting attrition may be more appropriately equated with other missing-data patterns rather than to complete data patterns. Returning to the illustration once more, an alternative set of identifying restrictions for the total nonresponse pattern could specify

$$f\left(Y_1 \mid (0,0)\right) = f\left(Y_1 \mid (1,0)\right)$$

$$f\left(Y_2 \mid (0,0)\right) = f\left(Y_2 \mid (0,1)\right)$$

The pattern-mixture model provides flexibility to make weaker assumptions about data relationships than those resulting from ignorability while maintaining the ability to estimate parameters needed for inference. Pattern-mixture models are not a panacea, however, as models must be specified not only for the conditional distribution of Y given the missing data pattern, but also for the relationships between parameters from different models. Specified models cannot be fully validated due to the missing data. Eltinge (2002) discusses considerations in evaluating methods for compensating for nonresponse.

E. Nonresponse Adjustment Approaches

As stated previously, statistical methods are commonly applied to compensate for nonresponse. Methods fall into two categories: (1) weight adjustment, in which sampling weights, based upon selection probability, for respondents are adjusted so as to account for the nonrespondents; or (2) imputation, in which values are assigned for the missing units, with appropriate sampling weights applied to all sample units, responding and imputed. Although weighting adjustment is the common method for compensating for unit nonresponse in cross-sectional surveys, imputation has desirable features for application with panel surveys. A number of discussions of common weighting and imputation methods are provided in the literature (see, e.g., Oh and Scheuren, 1983; Kalton and Kasprzyk, 1982; and, in the establishment survey setting, Hidiroglou, et al., 1995; Kovar and Whitridge, 1995).

1. Weighting Adjustments

Weight adjustments are discussed in detail in Oh and Scheuren (1983). The authors define two basic estimation approaches, poststratification and weighting class adjustment, which assume the population has been classified into subpopulations through either a response propensity or predicted means method. The choice of approach depends upon whether the population size is known for each subpopulation. Application of the two approaches is illustrated for the estimation of sample means.

For the poststratification approach, the estimated sample mean based upon the observed sample within each subpopulation, $\hat{\bar{Y}}_h$, is adjusted by the ratio of the post-stratum population size to the total population size

$$\hat{\bar{Y}}_{PS} = \sum_h \frac{N_h}{N} \hat{\bar{Y}}_h$$

while for the weighting class estimator, the adjustment is by the ratio of the weighting class sample size to the total sample size (which represents an estimate of the ratio of the weighting class population size to the total population size).

$$\hat{\bar{Y}}_{WC} = \sum_h \frac{n_h}{n} \hat{\bar{Y}}_h$$

Weight adjustments, although yielding appropriate estimates for means and totals of the population as well as for domains corresponding to weighting adjustment cells, are less efficient for estimates of population subgroups that do not correspond to weighting adjustment cells. Although nonresponse weighting adjustment can reduce bias in survey estimates, there is the potential for increased variance of the estimates through the creation of extreme weights or through increasing the variability of the weights beyond that intended by the sample design.

One reason extreme weights may result is due to the creation of a large number of adjustment cells due to cross-classification of a number of auxiliary variables. A method for controlling the generation of extreme weights and the resulting variance increase is raking ratio adjustment, or iterative proportional fitting (Deming and Stephan, 1940).

2. Imputation

An alternative approach to controlling both variability and bias resulting from nonresponse is imputation. Imputation involves the creation of appropriate values to represent those missing due to nonresponse. Imputation may also be used to create all values in the case of unit nonresponse. A key objective of imputation is to provide approximately unbiased estimates for the population of interest and domains of interest within the population. A variety of methods have been developed to provide imputed values for survey use. A model (either explicitly stated or implicitly assumed) relates the value for the unobserved unit to known information.

Kalton and Kasprzyk (1982) describe three desirable features of imputation: 1) imputation aims to reduce bias due to nonresponse; 2) imputation provides a complete data set for weighting and analysis; 3) results obtained from different analyses of a completed data set will be consistent. There are negative aspects to imputation as well. Imputation can result in increased bias. In addition, from a data use aspect, there is a risk analysts may view the completed data set as having been generated without nonresponse, and thereby understate the error when conducting analyses.

Kovar and Whitridge (1995) review approaches to imputation taken within business surveys. Imputation methods can be classified as deterministic or stochastic.

Deterministic methods yield imputed values that are uniquely determined given the sample of respondents. Stochastic methods, by contrast, yield imputed values that are subject to some degree of randomness. Often, the only difference between a deterministic and a stochastic method is the introduction of a random residual into the imputed value.

Following are categories of deterministic imputation methods employed within business surveys, as described by Kovar and Whitridge:

a. Mean imputation: Replaces missing values with the mean of the reported values within an imputation class. This method destroys distributions and multivariate relationships, and can perform poorly when nonresponse is not random. This method is equivalent to the weight adjustment approach, and assumes the following model.

$$Y_{ci} = \mu_c + \varepsilon_{ci}$$

b. Sequential hot-deck: Replaces data for a nonreporting unit with values from the last reporting unit preceding it in the data file. This method uses actual reported data for imputation, reasonably preserving distributions; however, care must be taken to minimize the frequency with which one respondent is imputed, to avoid effectively creating extreme weights. A critical issue is the choice of variables for formation of imputation classes and for sorting records within class.

c. Ratio and regression: Replaces missing values with corresponding ratio or regression predicted values, based upon some auxiliary variable(s). These methods are useful for imputing values for continuous variables, and perform well in cases of both random nonresponse and nonrandom but ignorable nonresponse. A critical step is obviously selection of the model and auxiliary variables. These approaches assume the following model.

$$Y_i = \mathbf{X}'\beta + \varepsilon_i$$

d. Nearest-neighbor: Replaces data for a nonreporting unit with values from a reporting unit of minimal distance (based upon some multivariate measure of the reported data) from the nonresponse unit. Like sequential hot-deck imputation, this method preserves multivariate relationships, but care must be taken to minimize the frequency with which one respondent is imputed.

Stochastic imputation can be represented by the general model

$$y_{mi} = b_{r0} + \sum_j b_{rj} x_{mij} + e_{mi}$$

where x_{mij} are the values of the auxiliary variables (indexed by j) for the i^{th} observation, b_{r0} and b_{rj} are the coefficients of a regression between y and x based on the responding units, and the e_{mi} are residuals chosen in a prespecified manner. The following categories of stochastic imputation method are commonly employed:

i. Random hot deck: Replaces data for a nonreporting unit with values from a randomly selected reporting unit from the data file. Selection may be either with or without replacement. This method better preserves distributions and limits multiple use of an individual donor record (especially with slection without replacement) more effectively than the sequential hot deck imputation method (Kalton and Kasprzyk, 1982).

ii. Regression with random residuals: Replaces missing values with corresponding regression predicted values, based upon some auxiliary variable(s), plus a residual.

As can be seen, the deterministic ratio and regression method fits the stochastic general model, with residuals set to zero. Correspondingly, the mean, sequential hot deck, and nearest neighbor deterministic methods could be applied as a stochastic method by adding random residuals.

If the data are MAR, stochastic imputation methods yield approximately unbiased estimates of distributions and element variances, while deterministic imputation methods tend to distort the shape of the distribution (Kalton and Kasprzyk, 1982). Mean and regression methods provide explicit models for the imputation; under the hot deck and nearest neighbor methods the imputation model is implicit.

Utilizing auxiliary information about the population, either through formation of imputation classes from which to estimate mean imputation values or directly as explanatory variables in a regression model, does provide the potential to reduce bias in survey estimates. Imputation also provides complete sample data sets, allowing more comprehensive population inferences than available with weight adjustments. Given the characteristics of establishment populations, with auxiliary data correlated with survey variables of interest commonly available for the universe, regression imputation models may be more desirable than mean imputation models for establishment surveys. When imputing for unit nonresponse under either a regression or mean imputation approach, a downside is the potential for attenuation as well as illogical or impossible combinations of variable values.

F. Mathematical Framework for Nonresponse in Panel Surveys

Consider a population of fixed size N. For each unit, $i\,(=1,\ldots,N)$, in the population, there is a variable of interest, Y_{ti}, for each reference period $t\ (=1,\ldots)$. The set of population values across reference periods can be represented by the column vector

$$\mathbf{Y}_{[Nt\times 1]} = \begin{bmatrix} \mathbf{Y}_{1[N\times 1]} \\ \vdots \\ \mathbf{Y}_{t[N\times 1]} \end{bmatrix} = \left[Y_{ti}\right]$$

with subvectors corresponding to the reference periods, and rows within each subvector corresponding to the units in the population.

It is assumed auxiliary information, possibly multivariate, about the population units is available, such that for each population unit there is a set of $Q\ (\geq 1)$ auxiliary variables (which may include values of the variable of interest, $\mathbf{Y}_{t*}$, for reference periods prior to t), such that the set of auxiliary variables can be represented by the matrix

$$\mathbf{X}_{[N\times Q]} = \left[X_{iq}\right]$$

In order to obtain estimates for the population statistics of interest, a panel survey is conducted, in which data are collected for each reference period from a sample, s, of fixed size $n\ \left(\leq N\right)$ selected from the population under some probability sample design, $p\left(s\right)$, such that the selection probability for unit i is π_i. The set of selection probabilities for the population can be represented by the vector $\boldsymbol{\pi}_{[N\times 1]} = \left[\pi_i\right]$.

The same set of sample units is surveyed across all months. Sample selection indicator $\delta_i = 1$ indicates unit i was selected, $\delta_i = 0$ indicates unit i was not selected. The population units may be ordered such that the set of sample selection indicators can be represented by the vector

$$\mathbf{I}_{[N\times 1]} = \begin{bmatrix} \mathbf{1}_n \\ \mathbf{0}_{(N-n)} \end{bmatrix}$$

Similarly, the set of population values can be partitioned into values for the sample units and values for the nonsampled units

$$\mathbf{Y}_{[Nt\times 1]} = \begin{bmatrix} \mathbf{Y}_{s[nt\times]} \\ \mathbf{Y}_{ns[(N-n)t\times 1]} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{s1[n\times 1]} \\ \mathbf{Y}_{ns1[(N-n)\times 1]} \\ \vdots \\ \mathbf{Y}_{st[n\times 1]} \\ \mathbf{Y}_{nst[(N-n)\times 1]} \end{bmatrix}$$

where $\mathbf{Y}_{st}$ is a subset of the full population vector $\mathbf{Y}_t$ for reference period t

corresponding to the sample units.

As a result of the survey environment, unit nonresponse occurs, yielding a reporting sample size for reference period t of $n_t\ (\leq n)$. Response indicators, r_{ti}, reflect the reporting status for sample unit i for reference period t. Response indicator $r_{ti} = 1$ signifies unit i reported reference period t data, while a response indicator $r_{ti} = 0$ signifies unit i did not report reference period t data. The set of response indicators for reference period t can be represented by the vector $\mathbf{R}_{st[n\times 1]} = \left[r_{ti} \right]$. The set of response indicators across all reference periods $t\ (=1,...,T)$ can be represented by the matrix

$$\mathbf{R}_{s\bullet[n\times t]} = \left[\mathbf{R}_{s1[n\times 1]} \quad \cdots \quad \mathbf{R}_{st[n\times 1]} \right]$$

The set of reported sample values can be represented by the matrix

$$\mathbf{Y}_{sR[n\times t]} = \left[r_{ti} Y_{ti} \right] = \left[Diag\left[r_{1i} \right] \mathbf{Y}_{s1} \quad \cdots \quad Diag\left[r_{ti} \right] \mathbf{Y}_{st} \right]$$

where $Diag\left[r_{ti} \right]$ is the diagonal $\left[n \times n \right]$ matrix with the response indicator r_{ti} as the i^{th} diagonal element.

Compensation for nonresponse in a panel survey then involves definition of working models specifying assumed distributions for the variable of interest in terms of the other available information (i.e., auxiliary variables, sample selection indicators, selection probabilities, and reporting status for the sample units across reference periods), then making use of available data $\left[\mathbf{Y}_{sR} : \mathbf{X} : \mathbf{I} : \boldsymbol{\pi} : \mathbf{R}_{s\bullet} \right]$ to derive estimates for the population, $\mathbf{Y}$ (i.e., deriving estimates for nonreporting units in the sample and for nonsample units in the population).

G. Nonresponse Adjustment for Panel Survey

The longitudinal nature of panel surveys brings the added dimension of time into consideration for nonresponse adjustment. Whereas cross-sectional surveys have only auxiliary information about nonreporting units available for use in nonresponse adjustment, panel surveys have available for the nonreporters values for the variable of interest from other reference periods (although often limited) which can be treated as additional auxiliary information for use in specifying the working models for the assumed distribution of the variable of interest for the current reference period.

Unit nonresponse adjustment for panel surveys generally follows one of three approaches – weight adjustment of the reported sample, imputation of the records for unit nonrespondents, or link relative estimation.

1. Weight Adjustments

Little and David (1983) proposed a method for adjusting for attrition in a panel survey, utilizing auxiliary information from the frame along with response information from periods prior to attrition. Sample weights are adjusted based on the regression of the response indicators for a wave and all available auxiliary information. This approach, however, only applies for strict attrition.

Kalton (1986) proposed a panel survey weight adjustment approach wherein nonrespondents and respondents for a time period are matched based on their reporting pattern for prior time periods. For example, letting 1 signify a response and 0 signify nonresponse, the approach would match the following sets of sample units

$$\begin{bmatrix}1 & 0 & 1 & 0\end{bmatrix}$$
$$\begin{bmatrix}1 & 0 & 1 & 1\end{bmatrix}$$

and weight up the time period 4 respondents to represent the time period 4 nonrespondents. This approach is rooted in the response propensity method, wherein sample units with the same prior reporting patterns are assumed to have similar distributions for the variables of interest. This method can be used to match reporters and nonreporters within adjustment cells defined on other auxiliary information. The underlying assumption is that nonresponse is not related to change in the variable of interest.

This approach becomes complex for surveys with large numbers of survey periods. Some patterns may have small numbers of respondents, so either large weights (and their corresponding impact on variance) must be accepted or reporting patterns must be collapsed. Special provisions must be made to handle more complex analyses, such as period-to-period change, as additional reporting patterns are no longer usable.

Kalton and Miller (1986) reported on the comparison of a weighting adjustment of respondents across all periods with a simple carry-over imputation (i.e., historical imputation) for a three-period panel survey simulated from the 1984 Panel of the Survey of Income and Program Participation (SIPP). Results showed carry-over imputation fared poorly, as it failed to represent changes over time. However, the applicability of this study is limited due to the small number of periods, the use of a carry-over imputation, and the restriction of the usable sample for weighting to units responding in all three periods.

Lepkowski (1989) provides an assessment of relative strengths of three weighting (total respondents, total respondents and strict attrition, all patterns), two imputation (carry-over, cross-period hot deck), and two combined (impute for patterns with limited numbers of missing periods and weight for all others, impute for selected non-attrition patterns to achieve attrition pattern and weight for all others using total respondent and strict attrition patterns) strategies in terms of five criteria

a. Practicality – ease of implementation and ease of use of subsequent data

b. Flexibility – ability of the procedure to handle multiple data types in a data record

c. Quality – ability of the procedure to predict the missing value correctly

d. Precision – accuracy of the resultant estimates

e. Preservation of relationships – maintains structure across variables

No strategy was deemed clearly superior. Weighting strategies were deemed preferable when the amount of period nonresponse is limited, and imputation strategies were deemed to have advantages when period nonresponse is substantial. Combined procedures were deemed to be worthy of consideration when the number of periods is large and weighting strategies falter. Several key points from the assessment which should be considered in looking at new approaches are: 1) incorporating as much prior information as possible into a weighting strategy provides the best opportunity to preserve relationships; 2) restricting a weighting strategy to respondents for all waves has a major negative impact on the precision of estimates; and 3) the validity of the working model is critical to the quality of the strategy.

Rizzo, et al. (1996), compare three approaches to weight adjustments for panel surveys: logistic regression; CHAID (Chi-square automatic interaction detector); and generalized raking. The logistic regression strategy sought to predict response rates within estimation cells, with three approaches used: full logistic regression prediction; prediction for small cells only with observed response rates used in large cells: and use of observed response rates in cells formed by collapsing smaller cells based on predicted response rates. The CHAID strategy created adjustment cells through application of two CHAID models – including seven most important predictor variables from the logistic regression model, and including all variables considered for the logistic regression model. The generalized raking strategy applied raking using marginal distributions for the predictor variables from the logistic regression model.

Comparisons were made using data from the 1987 panel for the SIPP. No substantive differences were found among the various methods. However, the authors found less correlation among alternative weights and the original SIPP weights, suggesting the choice of auxiliary variables is important. The authors also suggest using as many auxiliary variables correlated to response propensity as possible. This study looked at cross-sectional estimates and thus did not address the issue of change over time.

2. Imputation

Probably the most simplistic approach to imputation for panel surveys is historical imputation, as described by Kovar and Whitridge (1995):

Historical imputation uses values reported by the same unit on previous survey occasions. This method, while easily applied to unit nonresponse in panel surveys, will tend to attenuate size of trends and incidence of change, although variants adjust previous values by a measure of the trend. This method assumes the following model.

$$Y_{ti} = Y_{(t-1)i} + \varepsilon_{ti}$$

This model assumes there is no change in the value for a unit across reference periods, and thus is not a realistic working model given a key objective of a longitudinal survey is to measure change across time (as discussed in the study by Kalton and Miller, 1986). One area where such a model could be applicable is for surveys in which the variables of interest are categorical (e.g., labor force status) and strongly correlated over time for an individual unit.

Cross-wave hot deck imputation (see, e.g., Kalton, 1986), extends the stochastic hot-deck imputation method used for cross-sectional surveys. In this approach, nonrespondents for the current period are classified with respondents on the basis of reported information for a prior period when both responded. A donor respondent unit is randomly selected and that unit's current period information is imputed for the nonrespondent. However, given units are categorized in cells, information may still be lost.

Regression imputation for panel nonresponse (see e.g., Kalton, 1986) is also an extension from the cross-sectional environment, in this case of the cross-sectional regression imputation approach. Auxiliary variables include values from previous time periods, with the parameters estimated from the constant reporters.

Pfeffermann and Nathan (2002) proposed an extension of the proportional regression model, taking a time series approach. The time series model proposed for use in nonresponse imputation was of the form

$$Y_{ci(t)} = \mathbf{X}'_{ci(t)}\,\mathbf{b}_{(t)} + \mathbf{W}'_{c(t)}\,\mathbf{v}_{(t)} + \mathbf{W}'_{c(t)}\,\mathbf{u}_{c(t)} + e_{ci(t)}$$

where $\mathbf{X}_{ci(t)}$ is a p-dimensional vector of unit-level explanatory variables

$\mathbf{W}_{c(t)}$ is a q-dimensional vector of class-level explanatory variables

$\mathbf{b}_{(t)}$ and $\mathbf{v}_{(t)}$ are fixed vector coefficients

$\mathbf{u}_{c(t)}$ is a q-dimensional vector of class-level random effects, and

$e_{ci(t)}$ is a unit-level random error, with the unit-level and class-level random errors following independent first-order autoregressive models.

Based upon comparisons of bias and MSE for a simulated population the time series method was superior to both mean and simple regression imputation, and equivalent to augmented regression imputation.

3. Link Relative Estimation

An alternative to weighting and imputation sometimes employed for panel surveys of establishments is link relative estimation (Madow and Madow, 1978). Estimates of the relative change in the population total from one time period to the next are derived from the sample, and this estimated relative change is applied to the prior time period's estimated total. Although sample weights may be used in estimating the relative change, there is no adjustment of sampling weights or imputation for nonresponse.

Link relative estimation is a derivative of ratio estimation, the difference being a series of ratios are multiplied (or "linked") together to obtain the final ratio to be applied to the population value. In common practice, the ratios or links represent the relative period-to-period change for time periods beginning with that for which the population value is available through the current time period of interest.

For example, Madow and Madow (1978), define the link relative estimator for time t as

$$\hat{Y}_t = Y_0 \times LR_1 \times \ldots \times LR_t = Y_0 \prod_{t^*=1}^{t} LR_{t^*}$$

where Y_0 represents the population value at time 0

LR_{t*} represents the (known) estimated relative change (or link relative) from time period $t*-1$ to time period $t*$

Each link relative is derived on the basis of the reporting sample in time periods $t*-1$ and $t*$. Assuming formation of estimation cells, c, the estimator for the population total becomes.

$$\hat{Y}_t = \sum_c \left[\frac{\sum\limits_{i \in s_{t,(t-1)}} Y_{tci}}{\sum\limits_{i \in s_{t,(t-1)}} Y_{(t-1)ci}} \hat{Y}_{(t-1)c} \right] = \sum_c \left[\left(\prod_{t*=1}^{t} \frac{\sum\limits_{i \in s_{t*,(t*-1)}} Y_{t*ci}}{\sum\limits_{i \in s_{t*,(t*-1)}} Y_{(t*-1)ci}} \right) Y_{0c} \right] = \sum_c \left[\left(\prod_{t*=1}^{t} LR_{t*c} \right) Y_{0c} \right]$$

where $s_{t,(t-1)}$ represents the sample reporters common to reference periods t and $t-1$

The underlying model for the link relative estimator can be approximated by a proportional regression model with no intercept (Madow and Madow, 1978)

$$Y_{ti} = \beta Y_{(t-1)i} + \varepsilon_{ti}$$

$$\varepsilon_{ti} \sim \left(0, \sigma^2 Y_{(t-1)i} \right)$$

This proportional regression model has appeal for use in establishment surveys (it is used for the Current Employment Statistics survey), where inference is often made about the change or rate of change for the population. In that sense, this model can be thought of as a longitudinal analogue to the mean imputation model.

Although the link relative estimator uses prior information, it does not fully leverage the historical information about the relation between the nonrespondent and the respondents. The link relative also discards sample information from current time period in situation when reporters did not report data for the prior time period.

West, et al. (1989) examined the performance of four alternative proportional regression models in predicting actual values for employment data.

Model 1: $Y_{ti} = \alpha + \beta Y_{(t-1)i} + \varepsilon_{ti}$

Model 2: $Y_{ti} = \beta Y_{(t-1)i} + \varepsilon_{ti}$

Model 3: $Y_{ti} = \alpha + \beta \ln\left(Y_{(t-1)i}\right) + \varepsilon_{ti}$

Model 4: $Y_{ti} = \beta \ln\left(Y_{(t-1)i}\right) + \varepsilon_{ti}$

Errors were first assumed to have a simple variance structure, $\varepsilon_{ti} \sim \left(0, \sigma^2\right)$, then assumed to have a variance proportional to either the prior time period's level ($\varepsilon_{ti} \sim \left(0, \sigma^2 Y_{(t-1)i}\right)$, models 1-2) or the log of the prior time period's level ($\varepsilon_{ti} \sim \left(0, \sigma^2 \ln\left(Y_{(t-1)i}\right)\right)$, models 3-4). The authors found no one model superior to the others, but found Model 2 with error variance proportional to the prior time period's value (the same model described in Madow and Madow, 1978) robust, simple, and intuitively appealing. This study did not, however, examine more extensive use of prior information.

Previously, West (1983) had considered link relative type and regression type estimators utilizing information from the two prior time periods along with the basic one period link relative estimator. The one period link relative estimator again performed well when looking at estimates of both level and change, while the regression estimators tended to do poorly the longer the time period between the availability of the administrative data on population totals and the current time period.

H. Late Reporting

For many ongoing economic surveys, estimates are to be published soon after the reference period according to some prescribed processing schedule. The processing schedule requires completion of data collection as of some given cutoff date, resulting in unit nonresponse for the sample. Some of the unit nonresponse is temporal, as additional responses are obtained subsequent to the cutoff date. Given this late reporting, revised estimates for reference period t are often issued as part of processing for some fixed number of subsequent reference periods.

Revisions due to late reporting can be non-negligible. For the Current Employment Statistics survey, revisions between initial estimates and final estimates incorporating late reporters, while less than 0.1% at the national level, have varied by more than 1% for some industries (Copeland, 2003b). Monthly Retail Trades Survey revisions (which are due to both rotating sample and late reporting) have been less than 0.3% nationally, but as high as 5% for selected industries (Cantwell, et al., 1995). Revisions to the advanced sample estimates for the Statistics of Income Corporate Sample were as high as 11% for selected variables (Czajka and Hinkins, 1993).

To extend the mathematical framework for nonresponse to include late reporting, assume initial estimates for reference period t are based upon sample units reporting by a predefined initial cutoff date, d_t. Revised estimates for reference period t are issued concurrent with the initial estimates for each of the following K reference periods, with the revised estimates for reference period t incorporating all late reporting received to date. Late reporting for reference period t is thus accepted until a predefined final cutoff date, d_{t+K}, which also serves as the initial cutoff date for reference period $t+K$.

The cutoff date specific response indicator for sample unit i for reference period t, $r_{it|k}$, is defined as the response status relative to cutoff date, d_{t+k} $\left(0 \le k \le K\right)$. A cutoff date specific response indicator $r_{it|k} = 1$ signifies unit i reported reference period t data on or before cutoff date d_{t+k}, while a response indicator $r_{it|k} = 0$ signifies unit i had not reported reference period t data as of cutoff date d_{t+k}.

Note that:

1. $r_{it|k} = 1 \Rightarrow r_{it|k*} = 1, \ \left(k* \ge k\right)$; and
2. $r_{it|k*} = r_{it|K}$, $(k* \ge K)$, given the final cutoff date for reference period t is d_{t+K}

Response indicators for reference period t for unit i across cutoff dates may be summarized by the reporting status variable

$$\mathbf{X}_{ti} = \begin{pmatrix} X_{ti}^{PR} \\ X_{ti}^{LR} \\ X_{ti}^{NR} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

where the superscripts refer to preliminary reporting $\left(PR\right)$, late reporting $\left(LR\right)$, and nonresponse $\left(NR\right)$

$$X_{ti}^{PR} = \begin{cases} 1 \text{ if } r_{ti|0} = 1 \ \left(\text{PR for month } t\right) \\ 0 \text{ if } r_{ti|0} = 0 \end{cases}$$

$$X_{ti}^{LR} = \begin{cases} 1 \text{ if } r_{ti|0} = 0 \text{ and } r_{ti|K} = 1 \ \left(\text{LR for month } t\right) \\ 0 \text{ if } r_{ti|0} = 1 \text{ or } r_{ti|K} = 0 \end{cases}$$

$$X_{ti}^{NR} = \begin{cases} 1 \text{ if } r_{ti|K} = 0 \ \left(\text{NR for month } t\right) \\ 0 \text{ if } r_{ti|K} = 1 \end{cases}$$

The set of reporting status variables for all reference periods as of cutoff date d_t can be represented by the matrix

$$\mathbf{X}_{s\bullet|0_t[n\times t]} = \left[\mathbf{X}_{s1|K[n\times 1]} \quad \cdots \quad \mathbf{X}_{s(t-K)|K[n\times 1]} \quad \cdots \quad \mathbf{X}_{s(t-k)|k[n\times 1]} \quad \cdots \quad \mathbf{X}_{st|0[n\times 1]} \right]$$

Values of reporting status variables, $\mathbf{X}_{tci}$, only become known once a sample unit reports, or following the final cutoff date for reference period. However, preliminary estimates are based upon information known as of the initial cutoff date.

The accuracy of preliminary estimates for a reference period will depend upon (in addition to the sample design) nonreporting and late reporting size and patterns, the nature and magnitude of bias associated with the nonresponse, and the ability of the estimation methodology to eliminate, or at least reduce, these errors. The overall accuracy of survey estimates will depend upon the nature and magnitude of nonresponse bias. Failure of the estimation methodology to account adequately for nonresponse bias will result in potentially large benchmark revisions to final survey estimates. The accuracy of the preliminary estimates will also depend upon the nature and magnitude of any bias associated with late reporters. Failure of the estimation methodology to adequately account for late reporter bias will result in potentially large revisions to preliminary estimates.

Drew and Fuller (1981) explored the issue of estimation using information on late reporting relative to callbacks. The working assumption was that likelihood of response, q_k, depended upon some characteristic known for all sample units and was constant within characteristic for each contact. Drew and Fuller (1981) defined an estimator for the population mean based on R callbacks as

$$\hat{\bar{Y}} = \sum_{k=1}^{K} \hat{f}_k \bar{y}_k$$

where $\bar{y}_k$ is the sample mean for all respondents across all R callbacks

$$\hat{f}_k = \frac{\left[1-\left(1-\hat{q}_k\right)^R\right]^{-1} n_{.k}}{\left\{\sum_{k=1}^{K}\left[1-\left(1-\hat{q}_k\right)^R\right]^{-1} n_{.k}\right\}^{-1}}$$ is the estimated proportion of units with characteristic k

$\hat{q}_k$ is the solution to the polynomial equation

$$\sum_{r=1}^{R} n_{.k}^{-1} n_{rk}\left(1-rq_k\right) = \left[1-\left(1-q_k\right)^R\right]^{-1} Rq_k\left(1-q_k\right)^R$$

with k representing a characteristic and r representing a callback

This approach, while incorporating information about late reporting, assumes the distribution does not vary across callbacks within a characteristic, thereby requiring ignorability of both the late reporting and residual nonresponse mechanism.

Czajka, et al. (1992) proposed a response propensity approach to the problem of estimating corporate tax information from an advance sample of returns. Sample units within each design stratum were assigned to a propensity (of advanced reporting of tax information) class on the basis of auxiliary information, and weights were calculated within each propensity class and stratum using two methods

a. “Propensity stratification”

$$w_{jk}^{1} = \frac{\hat{N}_{jk}}{n_{jk}}$$

where $\hat{N}_{jk}$ is the estimated number of population units that would fall into propensity class k of stratum j

b. "Propensity weighting"

$$w_{jk}^{2} = w_{jk}^{2P} \frac{\hat{N}_{j\bullet}}{\sum_{k} w_{jk}^{2P} \times n_{jk}}$$

$$w_{jk}^{2P} = \sum_{i=1}^{n_{jk}} \frac{1/\left(1-\hat{p}_{ijk}\right)}{n_{jk}}$$

where $w_{jk}^{2P} = \sum_{i=1}^{n_{jk}} \frac{1/\left(1-\hat{p}_{ijk}\right)}{n_{jk}}$ is the preliminary weight

$\hat{p}_{ijk}$ is the predicted propensity for the i^{th} observation in propensity class k of stratum j

Results showed estimates from the propensity approach generally represented improvements (relative to the final estimates based upon the full – early and late – sample) over the existing approach to estimating from advanced reports (weighting based on design stratum only: $w_j = \frac{\hat{N}_j}{n_j}$). Results appeared consistent when looking at variables used in the propensity prediction and those not used.

The propensity approach could prove useful in application for panel surveys as well. The challenge would be to find predictors of response propensity/on time reporting propensity related to change over time, which is the key measure of interest.

Both the Current Employment Statistics survey (Bureau of Labor Statistics, 2004a) and Monthly Retail Trade Survey (Hogan, et al., 1997) generate preliminary estimates that are later revised to incorporate data from late reporters. In both situations, ignorability of both the late reporting and residual nonresponse mechanisms are assumed within estimation cells, with preliminary estimates based upon weighted link relative using early reporters

$$\hat{Y}_{t|t}=\sum_h\left[\frac{\sum_{i\in s_{t,(t-1)|t}} Y_{hit}}{\sum_{i\in s_{t,(t-1)|t}} Y_{hi(t-1)}}\hat{Y}_{h(t-1)}\right]$$

where $s_{t,(t-1)|t}$ represents the sample reporters common to months t and $t-1$ which reported by d_t, the preliminary cutoff date for month t, and final estimates based upon weighted link relative using both early and late reporters

$$\hat{Y}_{t|(t+k)}=\sum_h\left[\frac{\sum_{i\in s_{t,(t-1)|(t+k)}} Y_{hit}}{\sum_{i\in s_{t,(t-1)|(t+k)}} Y_{hi(t-1)}}\hat{Y}_{h(t-1)}\right]$$

where $s_{t,(t-1)|(t+k)}$ represents the sample reporters common to months t and $t-1$ which reported by d_{t+k}, the final cutoff date for month t.

Hogan, et al. (1997) examined the ability of simple linear models to improve the performance of advanced estimates for the Monthly Retail Trade Survey. Parameters were estimated based upon historical relationships between advanced and final estimates. Results were mixed. Approaches considered were fairly simplistic, however, with no attempt to incorporate other information which might have served to improve performance such as prior knowledge about late reporters or information about rates of change over time.

Rao, et al. (1989) proposed a time series approach, following the Kalman filter approach of Harvey (1981), for generating preliminary estimates based on early reporters. In the first approach, errors in the preliminary estimates are assumed to follow a stationary AR(1) process

$$\hat{Y}_t^* = \hat{Y}_t^P - \hat{Y}_t = \psi \hat{Y}_{(t-1)}^* + \zeta_t$$

and that the final estimates follow an AR(1) process

$$\hat{Y}_t = \phi \hat{Y}_{t-1} + \varepsilon_t$$

where ψ represents

$$\begin{pmatrix} \zeta_t \\ \varepsilon_t \end{pmatrix} \overset{ind}{\sim} N\left(0, \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right)$$

The second approach incorporates sampling errors about the final estimates

$$\hat{Y}_t = Y_t + u_t$$

where $u_t \sim iid\ N\left(0, \sigma_u^2\right)$, and assumes the population values follow an AR(1) process

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

The third approach extends the second approach to assume the errors may be correlated across time

$$\mathbf{u} \sim iid\ N\left(\mathbf{0}, \sum\nolimits_u\right).$$

These approaches were compared with the standard preliminary estimate approach in terms of estimating the final estimated level, $\hat{Y}_t$, the true level, Y_t, and the true period-to-period change, $Y_t - Y_{t-1}$, for profits data from quarterly surveys of industrial corporations conducted by Statistics Canada. Results indicate that, while the standard preliminary estimate is essentially the best predictor of the final estimate, $\hat{Y}_t$, the second time series approach performs better for estimating both the true level and the true period-to-period change. These approaches looked at the relationship between totals, rather than looking at the unit level. As a result, adjustment for late reporting was at an aggregate level, rather than by unit. Carrying the models down to lower levels could incorporate more information about relationships.

I. Discussion

Although the issue of compensating for nonresponse has been widely researched and addressed in cross-sectional surveys, less attention has been given to this issue for panel surveys. Many approaches can be seen as general extensions of cross-sectional methods. Other methods seek to model based on change from immediately prior period, and assume ignorable nonresponse for the existing period. This assumption may not be met in many applications and, as a result, estimates may not be accurately reflecting current levels and change from prior period.

A broad reporting pattern classification that accounts for both reporting status and timeliness of reporting may provide a structure for developing a pattern-mixture model to estimate growth rates without the assumption of ignorable nonresponse. Such a working model would seek to leverage prior information about nonreporters where available, thereby expanding from simpler models that only incorporate information about reporters.

Additionally, integration of nonresponse and late reporting models is needed to address the more appropriate view of the problem being faced in panel surveys with short publication deadlines. Such approaches may require a combination of modeling likelihood of both response and timeliness of reporting along with pattern-mixture models for estimating distributional properties.

Chapter III: Principal Motivating Example

The Bureau of Labor Statistics' (BLS) Current Employment Statistics (CES) survey is a monthly survey of establishments in the United States collecting information on employment, hours, and earnings. The primary statistics of interest for the CES survey are the total non-farm payroll employment in the U.S., and the change in total non-farm payroll employment from the prior month. CES estimates for these statistics are generated using data collected from a monthly panel survey, with a sample size over 300,000 establishments. In order to provide timely information, estimates are generated three to four weeks after the survey reference period. Estimates are revised each of the next two months to incorporate late reporting, and are subsequently revised on an annual basis to incorporate the most recent benchmark population information.

The reader is referred to Bureau of Labor Statistics (2001, 2004a, 2004b), upon which this chapter is based, for broader and more detailed descriptions of the CES survey. Appendix A contains a statistical formulation for a broader class of panel surveys, within which the CES survey is contained.

A. CES Sample Design and Data Collection

The population for the CES survey consists of over 8 million non-farm business establishments (defined as an economic unit which produces goods or services) in the United States. The population frame is derived from the BLS' ES-202 program, a federal/State cooperative between the BLS and State Employment Security Agencies (SESA's). The ES-202 program collects information on businesses covered by State unemployment insurance (UI) laws and Federal agencies covered by the

Unemployment Compensation for Federal Employees (UCFE) program. The main exclusions from this population are small agricultural employers and nonprofit organizations, and selected classes of workers (self-employed, domestic help, railroad workers, and State and local government elected officials).

The BLS recently completed a major redesign of the CES survey (Werking, 1997; Bureau of Labor Statistics, 2003), moving the survey from its historical quota sample design to a probability basis. The probability sample design was phased into published estimates over a four year period, with one or more major industry divisions transitioned from the quota sample to the probability sample each June, beginning in 2000 and completed in 2003, as shown in Table 1.

Table 1-CES Timing for Transition to Probability Sample

CES Timing for Transition to Probability Sample

Major Industry Division	**National series**	**State and area series**
Wholesale trade	June 2000	March 2001
Mining, Construction, Manufacturing	June 2001	March 2002
Transportation and public utilities; Finance, insurance, and real estate; Retail trade	June 2002	March 2003
Services	June 2003	March 2003

The new sample design is a stratified, simple random sample of establishments, clustered by UI account. Strata are defined by state, industry (based upon North American Industry Classification System (NAICS) categories), and employment size (defined as the maximum employment across the most recent 12 month period). Sampling rates for each stratum are determined through optimum allocation.

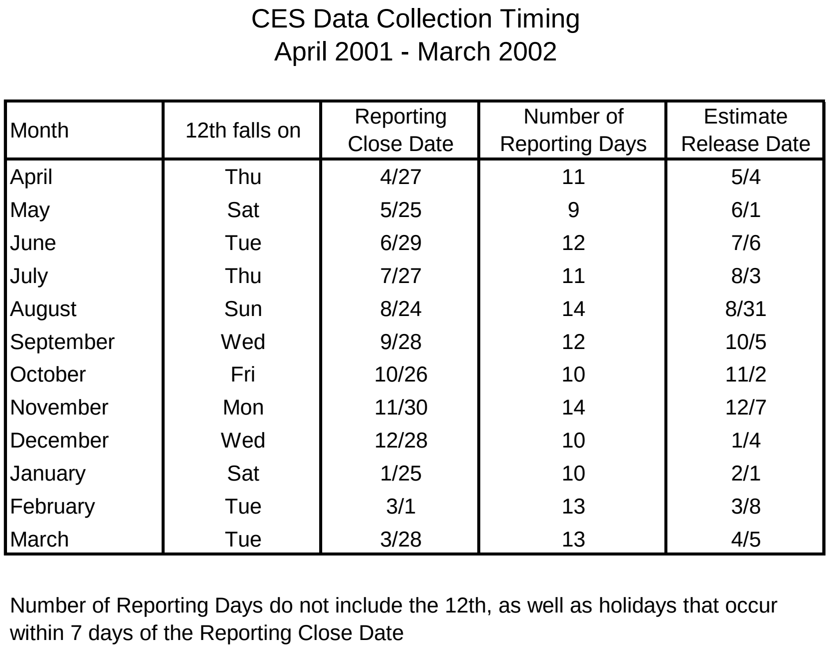
Sample selection is carried out on an annual basis, with the frame defined by the 1st quarter ES-202. Controlled selection is used to optimize overlap of sample establishments for both trending and operational efficiency. Sampling occurs late in the calendar year with new sample establishments sent to the field for data collection on a flow basis, to control workload; however new sample establishments are not immediately used in the estimation methodology. Sample replacement of the prior set of sample establishments with the new set of sample establishments in the estimation process occurs with the annual benchmarking process (described in section B). Thus there is approximately a two year lag between the time period used for frame development and implementation of the resulting sample into the CES estimates.

The BLS cooperates with the SESA's to collect the variables of interest from sample establishments. Respondents are asked to extract the requested data from payroll records. A variety of modes are used for data collection – touchtone data entry (TDE), computer-assisted telephone interviewing (CATI), mail, FAX, Electronic Data Interchange (EDI), magnetic tape, computer diskette, and World Wide Web (WWW). Regardless of mode, sample establishments are provided a "shuttle" form (BLS-790) reflecting the data to be provided for each month in the calendar year. The BLS-790 varies across industries, based on the specific information collected for each industry. (The BLS-790 for manufacturing is provided in Appendix B.1.)

The reference period for a given month is defined as the pay period that includes the 12th day of the month. The primary variable of interest for the CES survey is total employees, defined as persons on an establishment's payroll who received pay for any part of the pay period that includes the 12th day of the month. Other variables collected are women employees, and nonsupervisory/production/construction (depending upon the industry) employees along with their associated payroll, hours, and overtime hours.

All data must be reported within a two to three week period, the cutoff date depending upon the day of the week the 12th falls on and the number of days in the month, for inclusion in the initial published estimates for the month, which are generally released the first Friday of the following month. For example, data for July 2002 (for which the 12th was the second Thursday of the month) had to be reported by the cutoff date of July 27 (resulting in a reporting period of 11 calendar days from the 12th) to be included in the estimates published August 3. Table 2 contains information about CES collection timing for April 2001-March 2002.

Table 2-CES Data Collection Timing



CES Data Collection Timing
April 2001 - March 2002

Month	12th falls on	Reporting Close Date	Number of Reporting Days	Estimate Release Date
April	Thu	4/27	11	5/4
May	Sat	5/25	9	6/1
June	Tue	6/29	12	7/6
July	Thu	7/27	11	8/3
August	Sun	8/24	14	8/31
September	Wed	9/28	12	10/5
October	Fri	10/26	10	11/2
November	Mon	11/30	14	12/7
December	Wed	12/28	10	1/4
January	Sat	1/25	10	2/1
February	Tue	3/1	13	3/8
March	Tue	3/28	13	4/5

Number of Reporting Days do not include the 12th, as well as holidays that occur within 7 days of the Reporting Close Date

Not all sample establishments report by the cutoff date for the month. Additional responses are received after the close of the collection period for the month. Initial estimates for a given month (referred to as first closing estimates) are revised the subsequent two months, incorporating data from late reporters into the survey estimates. These revisions are referred to as second and third closing estimates.

Following is a standard classification of reporting status for sample establishments for a given month t, reflecting the CES collection methodology in terms of timing of reporting for current month reporters and, for current month nonreporters, prior reporting patterns.

1. Reporters
    a. Preliminary Reporters
        i. 1st Closing Reporters – sample establishments reporting data for the month prior to d_t, the cutoff date for processing preliminary estimates for month t
    b. Late Reporters
        i. 2nd Closing Reporters – sample establishments reporting data for the month after d_t but prior to d_{t+1}, the cutoff date for processing preliminary estimates for month $t+1$
        ii. 3rd Closing Reporters – sample establishments reporting data for the month after d_{t+1} but prior to d_{t+2}
2. Nonreporters – sample establishments not reporting data for month t
    a. Attritors – month t nonreporters which have reported for at least one prior month, but have not reported data for six or more consecutive months

b. Refusals – month t nonreporters which have not reported for any prior month

c. Episodic Nonreporters – all other month t nonreporters

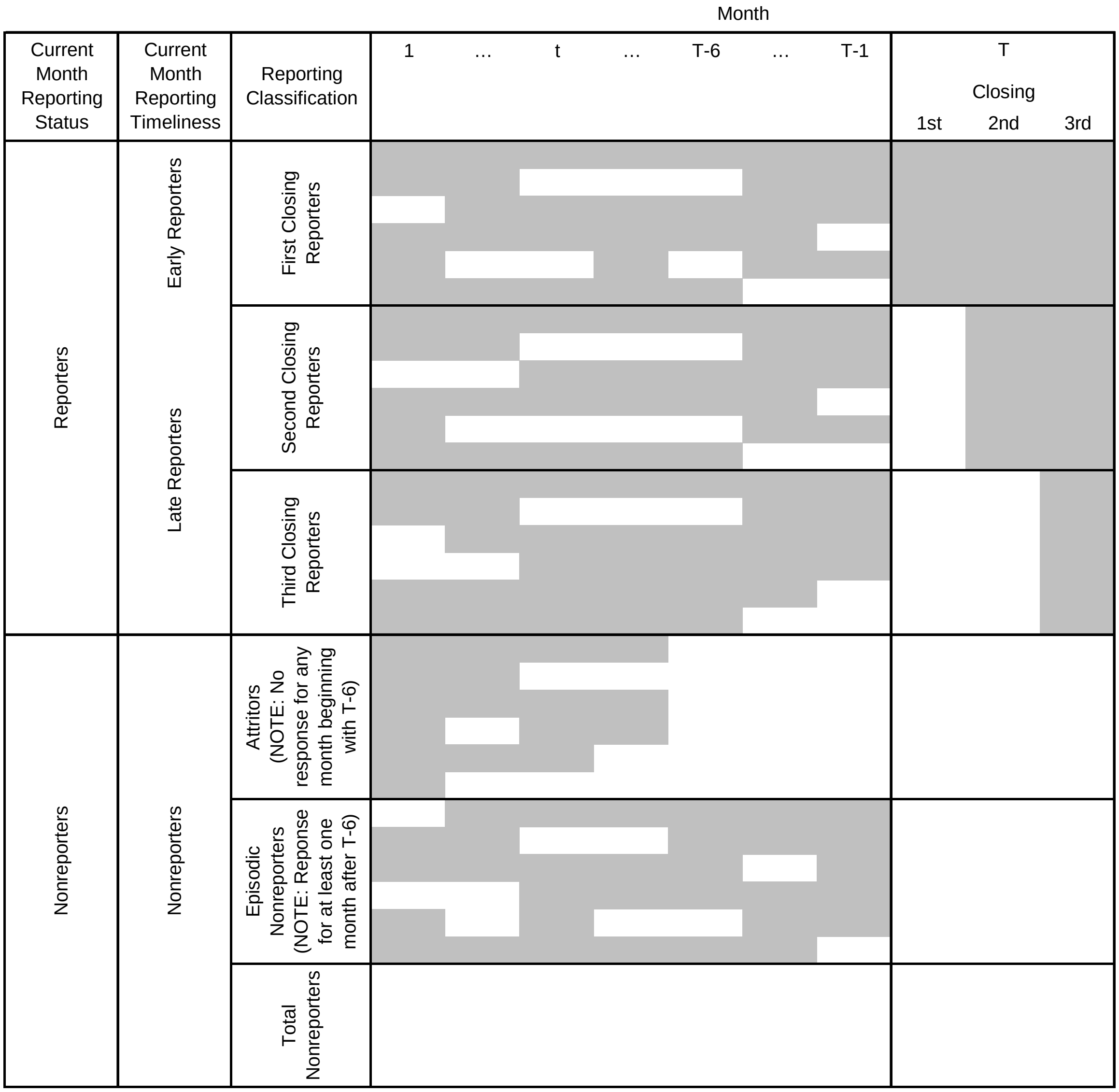
Figure 1 provides an illustration of these reporting patterns, with month t classification determined following subsequent months of data collection. All three nonreporter types (refusals, attritors, and episodic nonreporters) impact the overall accuracy of the CES estimates, regardless of closing. Late reporters (second closing reporters, third closing reporters) affect the accuracy of preliminary estimates only. The impact of late reporters on the preliminary estimates for month t can be assessed by examining the direction and magnitude of revisions between first and third closing estimates for month t. The impact of nonreporters on the final estimates can be assessed by examining the direction and magnitude of revisions between third closing estimates and benchmark data for the benchmark month (March).

On an annual basis, estimates are revised to reflect incorporation of ES-202 population data from March of the prior calendar year. These revisions are referred to as benchmark estimates. As part of benchmark estimation, data from late reporters beyond those included in the third closing estimates are included for selected months. In addition, sample replacement occurs during benchmark estimation.

Figure 1-CES Reporting Patterns

CES Reporting Patterns - Illustration for Month T

(Shaded area represents data reported for month, closing within month)



B. CES Estimation Methodology

CES survey estimates are generated through use of a weighted link relative estimator. This estimator uses a weighted sample trend within an estimation cell to move forward the prior month's estimate for that cell. The current CES weighted link relative estimator of all employees for a given revision, $k\ \left(=0,1,2\right)$, for month t is defined broadly as

$$\hat{Y}_t^{(k)} = \sum_{c=1}^{C}\left[\frac{\sum_{i\in s_{t,(t-1)|k}} w_{ci}y_{tci}}{\sum_{i\in s_{t,(t-1)|k}} w_{ci}y_{(t-1)ci}}\hat{Y}_{(t-1)c}^{(k+1)}\right] = \sum_{c=1}^{C}\left[LR_{t,(t-1)c}^{(k)}\hat{Y}_{(t-1)c}^{(k+1)}\right]$$

where y_{tci} represents total employment reported by sample establishment i in estimation cell c for month t

$c\left(=1,...,C\right)$ refers to estimation cell (defined by industry and, for selected industries, region)

w_{ci} represents the sampling weight for sample establishment i in estimation cell c

$s_{t,(t-1)|k}$ represents the set of sample units that reported data for both months t and $t-1$ as of the cutoff date for revision k [=0,1,2] of month t

$\hat{Y}_{(t-1)c}^{(k+1)}$ represents the prior month, $t-1$, weighted link relative estimate for estimation cell c based upon data reported as of the cutoff date for revision $k+1$ of month $t-1$ (which corresponds to revision k of month t)

$LR_{t,(t-1)c}^{(k)} = \frac{\sum_{i\in s_{t,(t-1)|k}} w_i y_{tci}}{\sum_{i\in s_{t,(t-1)|k}} w_i y_{(t-1)ci}}$ represents the link relative for month t based upon data reported as of the cutoff date for revision k of month t

As part of CES data processing, outliers are identified. Outliers are sample establishments reporting data yielding month-to-month changes that are viewed by survey analysts as abnormal or that report special reasons for the employment change from the prior month (e.g., strike). The CES estimator treats outliers as special cases, removing them from the sample included in the weighted link relative (and from the prior month's estimated population total) and then adding them in after the link relative is applied to the adjusted prior month's estimate. This outlier treatment can be represented as

$$\hat{Y}_t^{(k)} = \sum_{c=1}^{C} \left[\frac{\sum\limits_{i \in s_{t,(t-1)|k}, i \notin O_t} w_{ci} y_{tci}}{\sum\limits_{i \in s_{t,(t-1)|k}, i \notin O_t} w_{ci} y_{(t-1)ci}} \left(\hat{Y}_{(t-1)c}^{(k+1)} - \sum_{i \in s_{t,(t-1)|k}, i \in O_t} y_{(t-1)ci} \right) + \sum_{i \in s_{t,(t-1)|k}, i \in O_t} y_{tci} \right]$$

where O_t represents the set of outliers identified for month t

For the remainder of the chapter, outlier treatment is not included in the estimation formulae in the interest of space and of clearly conveying the core estimator. However, it should be remembered that the outlier treatment is part of the estimator. The treatment of outliers, although of interest relative to the overall accuracy of the weighted link relative estimator, is not included in the scope of this dissertation research.

More specifically, the weighted link relative estimators for all employees at each closing are

First closing (i.e., preliminary or revision 0) estimate of monthly employment, generated based upon data reported as of the first closing cutoff date for month t

$$\hat{Y}_t^{(0)} = \sum_{c=1}^{C} \left[\frac{\sum\limits_{i \in s_{t,(t-1)c|0}} w_{ci} y_{tci}}{\sum\limits_{i \in s_{t,(t-1)c|0}} w_{ci} y_{(t-1)ci}} \hat{Y}_{(t-1)c}^{(1)} \right] = \sum_{c=1}^{C} \left[LR_{t,(t-1)c}^{(0)} \hat{Y}_{(t-1)c}^{(1)} \right]$$

NOTE: the first closing estimates for month t use the second closing estimates for month $t-1$

Second closing (i.e., revision 1) estimate of monthly employment, generated based upon data reported as of the second closing cutoff date for month t

$$\hat{Y}_t^{(1)} = \sum_{c=1}^{C} \left[\frac{\sum\limits_{i \in s_{t,(t-1)|1}} w_{ci} y_{tci}}{\sum\limits_{i \in s_{t,(t-1)|1}} w_{ci} y_{(t-1)ci}} \hat{Y}_{(t-1)c}^{(2)} \right] = \sum_{c=1}^{C} \left[LR_{t,(t-1)c}^{(1)} \hat{Y}_{(t-1)c}^{(2)} \right]$$

Third closing (i.e., revision 2) estimate of monthly employment, generated based upon data reported as of the third closing cutoff date for month t

$$\hat{Y}_t^{(2)} = \sum_{c=1}^{C} \left[\frac{\sum\limits_{i \in s_{t,(t-1)|2}} w_{ci} y_{tci}}{\sum\limits_{i \in s_{t,(t-1)|2}} w_{ci} y_{(t-1)ci}} \hat{Y}_{(t-1)c}^{(2)} \right] = \sum_{c=1}^{C} \left[LR_{t,(t-1)c}^{(2)} \hat{Y}_{(t-1)c}^{(2)} \right]$$

NOTE: both the second and third closing estimates for month t use the same estimate of employment for month $t-1$, $\hat{Y}_{(t-1)c}^{(2)}$ (the third closing estimate for month $t-1$)

The corresponding estimators for month-to-month change in all employees for each closing are

First closing (i.e., preliminary or revision 0) estimate of month-to-month change in employment, generated based upon data reported as of the first closing cutoff date for month t

$$\Delta_{t,(t-1)}^{(0)} = \hat{Y}_t^{(0)} - \hat{Y}_{(t-1)}^{(1)}$$

NOTE: the first closing estimates of month-to-month change for month t use the second closing estimates for month $t-1$

Second closing (i.e., revision 1) estimate of month-to-month change in employment, generated based upon data reported as of the second closing cutoff date for month t

$$\Delta_{t,(t-1)}^{(1)} = \hat{Y}_t^{(1)} - \hat{Y}_{(t-1)}^{(2)}$$

Third closing (i.e., revision 2) estimate of month-to-month change in employment, generated based upon data reported as of the third closing cutoff date for month t

$$\Delta_{t,(t-1)}^{(2)} = \hat{Y}_t^{(2)} - \hat{Y}_{(t-1)}^{(2)}$$

NOTE: both the second and third closing estimates for month-to-month change in employment for month t use the same estimate of employment for month $t-1$, $\hat{Y}_{(t-1)c}^{(2)}$ (the third closing estimate for month $t-1$)

The CES estimator implicitly assumes the trend for both late reporters and nonreporters within an estimation cell is the same as for preliminary reporters that also reported data for the prior month. Although both late reporters and nonreporters contribute to variance and nonresponse bias present in the CES estimates, it is late reporting alone that drives revisions seen between preliminary and final estimates. The current CES estimator, however, assumes late reporting is a form of ignorable nonresponse and does not differentially adjust late reporters.

On an annual basis, as part of the generation of first closing estimates for January, administrative information on employment from the ES-202 program is incorporated into the CES estimates. This is accomplished by replacing estimated employment for the March of the prior year with the actual employment for that March from the ES-202 program. The replaced March is referred to as the “benchmark” month and the employment counts for the replaced March are referred to as the “benchmark” employment. Estimates for the 11 months prior to the benchmark month and for all months subsequent to the benchmark month are revised based upon the benchmark employment for the replaced March.

Benchmark estimates take several forms. First, benchmark estimates are generated for months subsequent to the new benchmark month (i.e., from April through October of the prior year). These estimates take the same form as previously. Estimates for April through October of the prior year utilize the original link relative derived as part of third closing processing for the month (i.e., do not incorporate data from the sample replacement nor from fourth closing reporters). Link relatives for November and December of the prior year (3rd and 2nd closing estimates, respectively) are derived using the new sample that was fielded beginning the prior year.

$Y_t^{(BM)} = \sum_{c=1}^{C} Y_{tc}^{BM}$, for the benchmark month (March of the preceding year)

$\hat{Y}_t^{(BM1)} = \sum_{c=1}^{C} \left[\frac{\sum_{i \in s_{t,(t-1)|2}(old)} w_{ci} y_{tci}}{\sum_{i \in s_{t,(t-1)|2}(old)} w_{ci} y_{(t-1)ci}} Y_{(t-1)c}^{(BM1)} \right]$, for the April of the preceding year (i.e., using the benchmark value for March of the preceding year to initialize the link relative estimation)

$\hat{Y}_t^{(BM1)} = \sum_{c=1}^{C} \left[\frac{\sum_{i \in s_{t,(t-1)|2}(old)} w_{ci} y_{tci}}{\sum_{i \in s_{t,(t-1)|2}(old)} w_{ci} y_{(t-1)ci}} \hat{Y}_{(t-1)c}^{(BM1)} \right]$, for May-October of the preceding year

$\hat{Y}_t^{(BM1)} = \hat{Y}_t^{(2)} = \sum_{c=1}^{C} \left[\frac{\sum_{i \in s_{t,(t-1)|2}(new)} w_{ci} y_{tci}}{\sum_{i \in s_{t,(t-1)|2}(new)} w_{ci} y_{(t-1)ci}} \hat{Y}_{(t-1)c}^{(BM1)} \right]$, for November of the prior year

$$\hat{Y}_t^{(BM1)} = \hat{Y}_t^{(1)} = \sum_{c=1}^{C}\left[\frac{\sum\limits_{i \in s_{t,(t-1)|1}(new)} w_{ci} y_{tci}}{\sum\limits_{i \in s_{t,(t-1)|1}(new)} w_{ci} y_{(t-1)ci}} \hat{Y}_{(t-1)c}^{(BM1)}\right]$$, for December of the prior year

where $s_{t,(t-1)|2}(old)$ represents the set of outgoing sample units that reported data for both months t and $t-1$ as of the cutoff date for the 2^{nd} revision

$s_{t,(t-1)|k}(new)$ represents the set of incoming sample units that reported data for both months t and $t-1$ as of the cutoff date for the k^{th} revision

Second, benchmark estimates are generated for months prior to the new benchmark month, but subsequent to the previous benchmark month (i.e, from April two years prior through February of the prior year). The estimates for these months are the prior benchmark estimate for the month (generated the prior year) adjusted for the change in March to March employment levels based on the new benchmark data.

$$\hat{Y}_t^{(BM2)} = \sum_{c=1}^{C}\left[\left(Y_{t_{B_1}c} - Y_{t_{B_0}c}\right)\frac{t - t_{B_0}}{12} + \hat{Y}_{tc}^{(BM2)}\right]$$, for April two years preceding through February of the preceding year

where t_{B_1} represents the benchmark month from the preceding year

t_{B_0} represents the benchmark month from the two years preceding

Publication schedule for the CES survey is illustrated in Table 3 from two perspectives – by publication month and for a given reference month across publication months. Refer to Appendix B.2 for a chart indicating estimate revision schedules and data used.)

Table 3-CES Publication Schedule

CES Publication Schedule

Calendar Month	Published Estimates				
	1st Closing	2nd Closing	3rd Closing	First Benchmark Revision	Second Benchmark Revision
Nov '03	$Y_{Oct'03}^{(0)}$	$Y_{Sep'03}^{(1)}$	$Y_{Aug'03}^{(2)}$		
Dec '03	$Y_{Nov'03}^{(0)}$	$Y_{Oct'03}^{(1)}$	$Y_{Sep'03}^{(2)}$		
Jan '04	$Y_{Dec'03}^{(0)}$	$Y_{Nov'03}^{(1)}$	$Y_{Oct'03}^{(2)}$		
Feb '04	$Y_{Jan'04}^{(0)}$	$Y_{Dec'03}^{(1)}\left(=Y_{Dec'03}^{(BM1)}\right)$	$Y_{Nov'03}^{(2)}\left(=Y_{Nov'03}^{(BM1)}\right)$	$Y_{Apr'03}^{(BM1)},\ldots,Y_{Oct'03}^{(BM1)}$	$Y_{Apr'02}^{(BM2)},\ldots,Y_{Feb'03}^{(BM2)}$

June 2003 Publication Schedule	
Calendar Month	June '03 Estimate Published
Jul '03	$Y_{Jun'03}^{(0)}$
Aug '03	$Y_{Jun'03}^{(1)}$
Sep '03	$Y_{Jun'03}^{(2)}$
Feb '04	$Y_{Jun'03}^{(BM1)}$
Feb '05	$Y_{Jun'03}^{(BM2)}$

The CES estimator can thus be viewed as being initialized at month $t = 0$ by using the most recently available March data from the BLS' ES-202. The preliminary estimator (and, correspondingly, revised estimators) can be rewritten as the product of link relatives

$$\hat{Y}_t^{(0)} = \sum_{c=1}^{C}\left[\left\{LR_{t,(t-1)c}^{(0)}LR_{(t-1),(t-2)c}^{(1)}\prod_{t^*=1}^{t-2}LR_{t^*,(t^*-1)c}^{(2)}\right\}Y_{0c}\right]$$

where Y_{0c} represents the benchmark total employment from the ES-202.

The first two terms in the equation for $\hat{Y}_t^{(0)}$ represent the 1st and 2nd closing link relatives for months t and $t-1$, respectively. These terms will change as part of 2nd and 3rd closing estimation for Y_t. All other terms represent the 3rd closing link relatives for their respective months, which will not change.

Both monthly level and month-to-month change estimates from the CES survey are of interest to data users. Indirect measures of the accuracy of CES estimates are visible through the revision and benchmark process. Revisions from first to third closing revisions for all months except November and December are solely due to the effect of late reporting (November third closing and December second closing estimates also reflect the incorporation of the new benchmark data), while revisions from third closing to the benchmark for March are the result of the combined effects of sampling, nonresponse, and measurement error.

CES survey estimates are also adjusted to account for business births (new establishments) and deaths (closed establishments). Business deaths are excluded from the CES weighted link relative estimator; however, the prior month employment for such establishments is implicitly carried forward to the current month, thus overstating employment. This overstatement is offset by an understatement of the employment due to business births. As employment associated with business births will not equal the carried forward employment associated with business births, the residual employment due to the net effect of business births and deaths is estimated through use of a model-based approach.

CES survey estimates are seasonally adjusted to stabilize trends and enable better estimation of month-to-month changes in employment. Seasonal factors are calculated twice a year using multiplicative models in X-12 ARIMA, and revisions are made annually concurrent with the benchmark revision process.

Variance estimation for the CES survey is carried out using Fay’s method for variance estimation under balanced repeated replication (Judkins, 1990). A total of

80 balanced half-samples were selected. Using the Fay method, the CES variance estimator applies weights of 0.5 and 1.5 for the half-samples within a replicate, rather than the normal weights or 0 and 2. Thus

$$wt_{\alpha i} = 1 + 0.5 * I_{\alpha i}$$

where $I_{\alpha i} = (1, -1)$ represents the indicator assigned to distinguish which half-sample unit i belong to replicate α $(=1, \dots, 80)$

The estimate for the α^{th} replicate can then be represented as

$$\hat{Y}_{\alpha tc} = \frac{\sum\limits_{i \in s_{t,(t-1)|k}} wt_{\alpha i} w_{ci} Y_{tci}}{\sum\limits_{i \in s_{t,(t-1)|k}} wt_{\alpha i} w_i Y_{(t-1)i}} \hat{Y}_{\alpha(t-1)c}$$

The variance estimate is derived as an adjusted mean squared error

$$v\left(\hat{Y}_{tc}\right) = \frac{\sum\limits_{\alpha} \left(\hat{Y}_{\alpha tc} - \hat{Y}_{tc}\right)^2}{80(0.5)^2}$$

Variance estimates represent only sampling variance and do not reflect nonsampling errors, such as measurement error and nonresponse bias. Overall performance of the estimates is measured in terms of the size of the benchmark revisions (difference between third closing estimate and benchmark data for March).

C. Analysis Data Files

Analyses carried out as part of this dissertation research utilized CES sample data for the period January 2000 through December 2002, along with ES-202 population totals for March 2001 and 2002, for establishments from the four industries—Construction, Manufacturing, Mining, and Wholesale Trade—which had transitioned to a probability sample design as of March 2001. Data preparation was carried out using SAS v8.2.

1. Sample Data File

Sample establishments included in the analysis were those selected for the 2000 sample replacement which had reported employment data prior to 3rd closing in at least one month in the period January 2000 through December 2002. Given the controlled selection utilized for the CES survey, the majority of the 2000 sample had been previously selected and thus already in the field at the beginning of 2001, while newly selected establishments not previously in sample were sent to the field during 2001. (For the analysis datafile, 71.7% of establishments reported data in January 2000, and 90.8% of establishments had reported data prior to the start of the analysis period.) The 2000 sample was officially utilized for CES estimates effective May 2002, as part of the March 2001 benchmark revision. As part of that benchmark revision, estimates back to October 2001 were revised to utilize the 2000 sample.

A total of 60,944 sample establishments met the inclusion criteria. The datafile of included sample establishments was created as follows.

a. Reporters from the 2000 sample were extracted from the CES microdata files for January 2000 through December 2002 to create an initial datafile of CES reporters. The following data items were extracted from the microdata files: establishment CES identification number; data month and year; sample year; reported employment; closing for which data were reported; and class flag and explanation code (used for identifying atypicals and unusables).

This datafile was restructured to the unique establishment CES identification number level, with other data items reformatted to include a data month indicator. Data months for which no record existed for the establishment CES identification number on the CES microdata file were flagged as nonreporting.

A review of the data indicated atypical flags were not always indicated where needed, due to data preparation operations prior to transition to the 2000 sample for CES estimation. Following consultation with CES support staff at BLS, a custom process for flagging atypicals not previously identified was undertaken as part of data preparation. The custom process identified atypicals as those establishments for which month-to-month employment change was both greater than 100 and greater than 1.5 times the average of the current and prior months' reported employment. The number of establishments identified as atypical in any month never exceeded 45, and averaged 20 for the analysis period, representing 0.03% of the 60,944 establishments on the analysis datafile.

The SAS code used to read the CES microdata file and create an initial datafile of 253,972 CES reporters is provided in Appendix C.1.

b. Establishments in the 2000 CES sample for the industries of interest were extracted from the CES cross-walk file, which contains both design and other auxiliary information for establishments selected for the CES sample. The following data items were extracted for use in the analysis: establishment longitudinal database (LDB), Unemployment Insurance (UI), and RUN (reporting unit number) sample reporting number and reporting-with number; and NAICS industry code.

Sample reporting-with numbers are intended to link sample establishments that are reported together on one file. The establishment reporting data for one or more sample establishments is identified by the sample reporting-with number. The CES cross-walk file was segmented into a parent file (those records for which sample reporting number equaled sample reporting-with number) and a child file (those records for which sample reporting number did not equal sample reporting-with number).

The initial datafile of CES reporters was merged with the file of establishments from the CES cross-walk file, first by matching CES identification number from the CES datafile to sample reporting number from the parent CES cross-walk file, then by matching CES identification number from the unmatched CES datafile to sample reporting with number from the child CES cross-walk file. The full set of 253,918 matched records (99.98% of total records on initial CES datafile) was used to create a revised CES datafile.

The CES cross-walk file was used to append NAICS codes to the CES microdata file, and to pick up UI and RUN identification numbers for use in merging with other data files. The revised datafile of CES reporting was restricted to records with a NAICS code in Mining (113300 – 113399, 210000 – 219999), Construction (230000 – 309999), Manufacturing (310000 – 419999), and Wholesale Trade (420000 – 439999). This yielded a datafile consisting of 60,944 records.

The SAS code used to extract data from the CES cross-walk-file and merge with the initial datafile of CES reporters is provided in Appendix C.2 – C.4.

c. Information on length of pay period was obtained from the CES registry file for August 2001. The CES registry file contains information relative to sample recruitment and data collection. Registry files are maintained at the state level, however, and information is not consistently updated or maintained. As a result, only length of pay period was obtained from the CES registry file. August 2001 was used as this roughly corresponded to when fielding of the 2000 CES sample was complete. A total of 54,410 of the sample records (89.28%) on the CES datafile were matched to a record on the CES registry file. This subset formed the basis for parameter estimation under Bayes' models, while the full dataset was used for post-stratification and estimation.

The SAS code used to extract information from the August 2001 CES registry file and append it to the CES datafile is provided in Appendix C.5.

d. Selected sample design information (sample design size class, selection weight) for CES sample establishments is contained on the CES random group file for a given year's sample. The revised datafile of CES reporting was merged with the 2000 CES random group file on the basis of state and UI. number. The full set of matched records was used to update the revised CES datafile, appending design size class and selection weight. A total of 60,926 records (99.98%) on the CES data file were matched to the CES random group file.

The SAS code used to extract data from the CES random group file and merge with the revised CES datafile is provided in Appendix C.6.

Table 4 contains information on record counts for each step in the process of creating the CES datafile.

Table 4-CES Data File Record Counts

Preparation of CES Data File

Records on CES microdata file	253,972	
Includes all industries		
Report by 3rd closing at least one month in 1/00 - 12/02		
Matched to cross-walk file	253,918	99.98%
In one of four industries of interest	60,944	24.00%
Records used in post-stratification, estimation		
Matched to CES random group file	60,926	99.97%
Marched to August '01 CES Registry File	54,410	89.28%
Records used in parameter estimation under Bayes' models		

Table 5provides distribution information for selected characteristics for the CES microdata file.

Table 5-CES Datafile Distribution by Selected Characteristics

CES Microdata File
Distribution by Selected Characteristics

	Total	%
Total Establishments	60,944	100.0%
Industry		
Construction	16,739	27.5%
Manufacturing	29,742	48.8%
Mining	2,358	3.9%
Wholesale Trade	12,105	19.9%
Design Size Class		
<10	8,687	14.3%
10-19	5,429	8.9%
20-49	7,966	13.1%
50-99	7,232	11.9%
100-249	13,526	22.2%
250-499	6,280	10.3%
500-999	4,110	6.7%
1000+	7,696	12.6%
Missing	18	0.0%
Length of Pay Period		
Weekly	34,702	56.9%
Bi-Weekly	11,184	18.4%
Semi-Monthly	6,100	10.0%
Monthly	2,423	4.0%
Missing	6,535	10.7%

Table 6 provides reporting status counts and percentages by month for the CES microdata file. Reporting status counts are provided relative to total sample units (PR+LR+NR), reporting sample units (PR+LR), and non-preliminary reporters (LR+NR).

Table 6-CES Datafile Distribution by Reporting Status

CES Microdata File
Distribution by Reporting Status
April 2001 - March 2002

	Preliminary Reporters			Late Reporters			Nonresponders	
Month	Tot	% (PR+LR+NR)	% (PR+LR)	Tot	% (PR+LR+NR)	% (LR+NR)	Tot	% (PR+LR+NR)
Apr '01	33,065	60.8%	82.3%	7,125	13.1%	36.7%	12,264	22.5%
May '01	30,421	55.9%	76.6%	9,314	17.1%	41.9%	12,915	23.7%
Jun '01	32,043	58.9%	80.7%	7,685	14.1%	36.9%	13,157	24.2%
Jul '01	32,413	59.6%	80.7%	7,743	14.2%	36.0%	13,768	25.3%
Aug '01	33,817	62.2%	83.1%	6,864	12.6%	32.7%	14,117	25.9%
Sep '01	34,644	63.7%	83.3%	6,930	12.7%	32.3%	14,535	26.7%
Oct '01	33,295	61.2%	82.1%	7,280	13.4%	30.8%	16,346	30.0%
Nov '01	31,237	57.4%	75.8%	9,964	18.3%	37.9%	16,319	30.0%
Dec '01	31,150	57.3%	74.7%	10,542	19.4%	39.3%	16,278	29.9%
Jan '02	30,823	56.6%	76.5%	9,444	17.4%	34.4%	17,985	33.1%
Feb '01	33,946	62.4%	83.3%	6,821	12.5%	27.8%	17,675	32.5%
Mar '02	34,107	62.7%	83.6%	6,687	12.3%	27.1%	17,963	33.0%

For purposes of estimation for both parameter estimation for the employment growth model and the reporting status model, and for post-stratification and link relative estimation, the CES microdata file was restructured to create records at the sample establishment by month level, with information on the CES microdata file reformatted for ease of processing. The SAS code used to create the two analysis files is provided in Appendix C.7.

2. Benchmark Data

Population employment totals for March of 2000, 2001, and 2002 for the industries of interest were derived from BLS' Longitudinal Database (LDB), which is the basis for the ES-202. All establishments within the industries of interest as of 1st quarter 2000 were extracted from the LDB, along with reported employment for

March of 2000, 2001, and 2002. Employment data were summed to the industry level to obtain benchmark figures for each month. The SAS code used to summarize LDB data is provided in Appendix C.8.

Chapter IV: CES Reporting Pattern Profile

The CES survey is subject to late reporting and nonresponse, which are the result of a combination of respondent, operational, and environmental factors. Understanding the reporting dynamics for the survey can not only identify opportunities for improving response rates, but may also suggest working models for predicting response status and for imputing for missing data due to late reporting and nonresponse.

Reporting patterns are of interest for two reasons. First, the extent and recency of information available for use in estimation varies across reporting patterns. Second, distributional properties may differ among the patterns. Both should be taken into account when specifying factors for the underlying working model used for imputation.

As discussed in Chapter II, late reporting and nonresponse can adversely affect the quality of survey estimates. For panel surveys, the patterns of late reporting and nonresponse across time are of interest as well as their levels. Prior to profiling CES survey reporting patterns, a new taxonomy for classifying reporting patterns in panel surveys, extending prior work in this area, is developed. This taxonomy is then tied into the CES survey classifications of reporting status to define an approach for looking at CES reporting patterns.

CES reporting patterns were profiled using data from January 2000 through December 2002 for four industries (Construction, Manufacturing, Mining, and Wholesale Trade), encompassing a total of 60,944 sample establishments reporting data for at least one month in the period (regardless of timeliness). Reporting patterns were profiled in several ways. First an overview of CES reporting patterns is provided relative to the new reporting pattern taxonomy and CES reporting status categories, and then month-to-month reporting patterns are provided relative to the structure of the CES weighted link relative estimator and the interest in developing a model to allow imputation for missing data.

Based upon information gleaned from the profile of the CES survey reporting patterns, a model is developed for predicting reporting status for sample units not reporting for 1st closing. The adequacy of the model is evaluated on the basis of comparison to actual reporting status at the aggregate level.

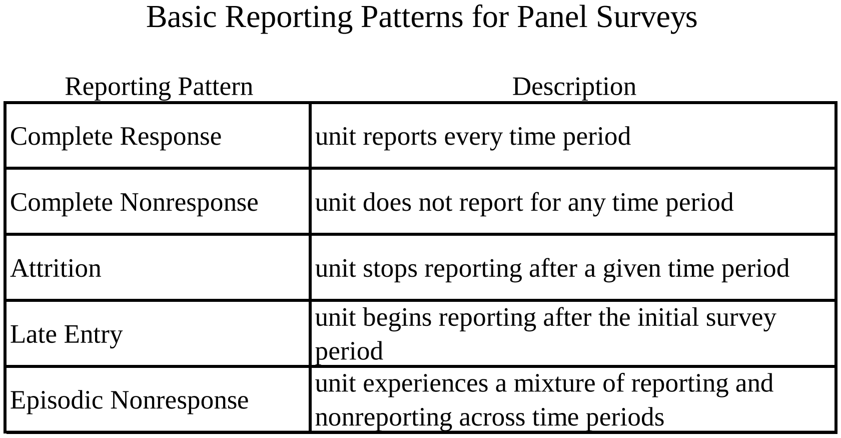
A. A New Taxonomy for Panel Reporting Patterns

Survey nonresponse is frequently classified on the basis of reason for nonresponse. Panel surveys add another dimension to the response mechanism, that being response status by survey period. Surveys that publish revised estimates offer yet another dimension to the response mechanism, that being timeliness of reporting.

As discussed in Chapter II, existing taxonomies for nonreporting patterns could be refined to reflect more completely the nature of reporting patterns. Clarifying distinctions among patterns could prove useful for both response improvement efforts, by providing greater granularity for nonresponse analyses, and development of nonresponse adjustment methods, as distributional properties could differ among patterns.

Reporting patterns can be categorized into five basic types, as shown in Table 7.

Table 7-Basic Reporting Patterns for Panel Surveys



Basic Reporting Patterns for Panel Surveys

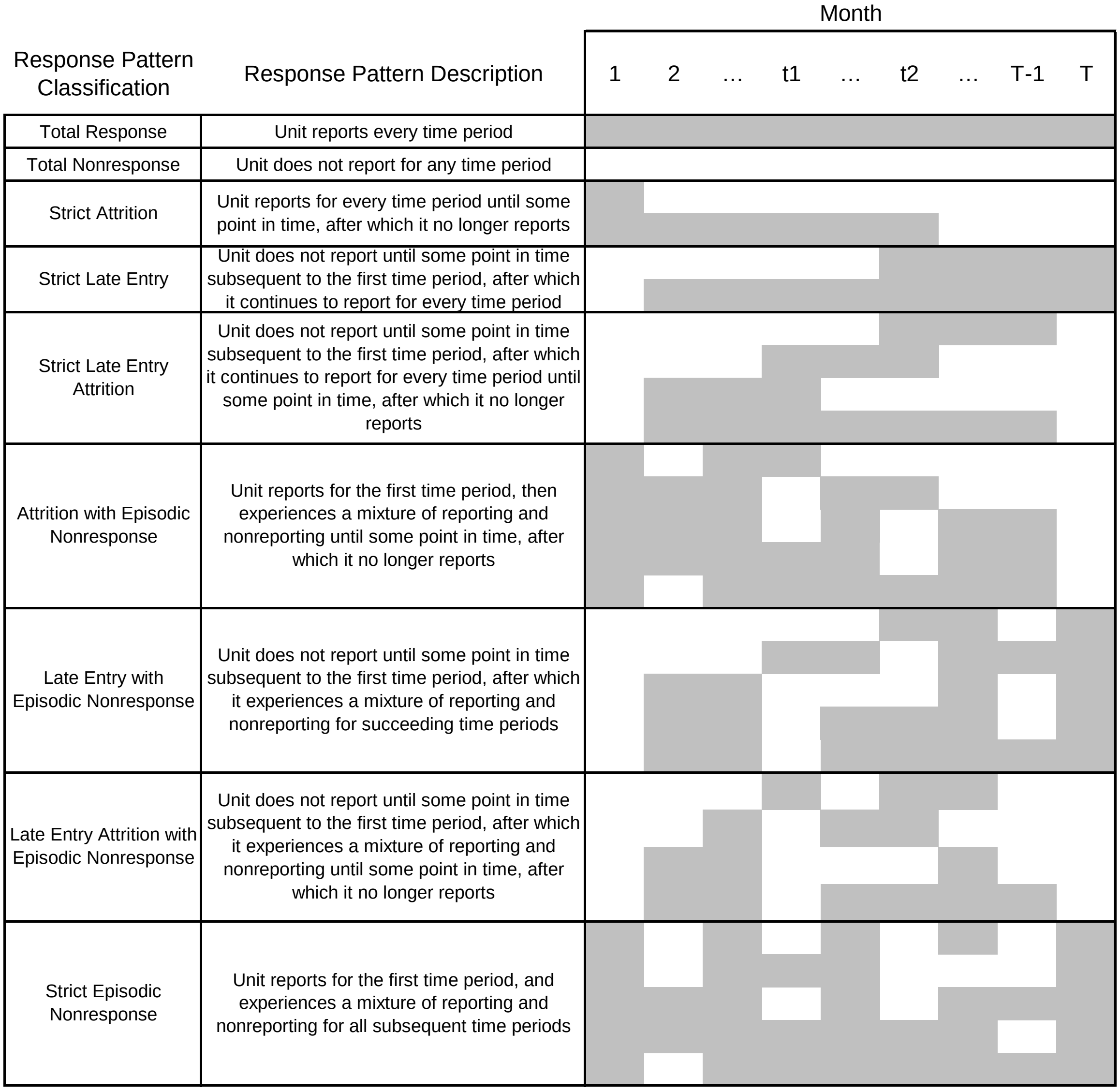
Reporting Pattern	Description
Complete Response	unit reports every time period
Complete Nonresponse	unit does not report for any time period
Attrition	unit stops reporting after a given time period
Late Entry	unit begins reporting after the initial survey period
Episodic Nonresponse	unit experiences a mixture of reporting and nonreporting across time periods

An expanded and refined set of reporting patterns for panel surveys can be defined by mixtures of the basic reporting patterns. Reporting patterns defined by only one basic pattern may be thought of as first order reporting patterns, while other reporting patterns (based upon a combination of basic patterns) may be thought of as interactions of reporting patterns. This taxonomy for reporting patterns, along with illustrations, is provided in Figure 2. Note that classification of a sample unit in terms of a reporting pattern is temporary, unless the survey has ended and there will be no further time periods for which data will be collected.

Figure 2-Reporting Pattern Illustrations

Response Pattern Illustrations

Shaded area represents data reported for month


Month
1
2
…
t1
…
t2
…
T-1
T

Response Pattern Classification	Response Pattern Description
Total Response	Unit reports every time period
Total Nonresponse	Unit does not report for any time period
Strict Attrition	Unit reports for every time period until some point in time, after which it no longer reports
Strict Late Entry	Unit does not report until some point in time subsequent to the first time period, after which it continues to report for every time period
Strict Late Entry Attrition	Unit does not report until some point in time subsequent to the first time period, after which it continues to report for every time period until some point in time, after which it no longer reports
Attrition with Episodic Nonresponse	Unit reports for the first time period, then experiences a mixture of reporting and nonreporting until some point in time, after which it no longer reports
Late Entry with Episodic Nonresponse	Unit does not report until some point in time subsequent to the first time period, after which it experiences a mixture of reporting and nonreporting for succeeding time periods
Late Entry Attrition with Episodic Nonresponse	Unit does not report until some point in time subsequent to the first time period, after which it experiences a mixture of reporting and nonreporting until some point in time, after which it no longer reports
Strict Episodic Nonresponse	Unit reports for the first time period, and experiences a mixture of reporting and nonreporting for all subsequent time periods

For a survey such as the CES survey, in which revised estimates for a given month are generated, late reporting adds another dimension to reporting patterns, as illustrated in Figure 3. In order for a sample unit to be utilized in the first-closing link relative, it must have reported for the prior month (whether preliminary or late) as well as have been a preliminary reporter for the current month.

Figure 3-Timeliness Pattern Illustrations

Timeliness Pattern Illustrations

Shaded area represents data reported on-time for month
Dotted area represents late reported data for month

		Month		
Response Pattern Classification	Timeliness Classification	T-1	T	Use
Current, Prior Month Reporter	On-Time both months	▨	▨	Preliminary
	On-time current month only	░	▨	Preliminary
	On-time prior month only	▨	░	Final
	Late both months	░	░	Final
Current, Prior Month Nonreporter	N/a			No
Prior Month Only Reporter	On-time	▨		No
	Late	░		No
Current Month Only Reporter	On-time		▨	No
	Late		░	No

B. CES Reporting Patterns Relative to Taxonomy

The focus of this profile is on the dynamic portion of CES survey nonreporting – attrition, and episodic nonreporting. Complete nonresponse, while contributing to the overall nonresponse impact, is less tractable in terms of a nonresponse adjustment strategy due to the lack of any reported data. Late reporting is discussed in the next section. Portions of the results presented in this section have been described elsewhere (Copeland 2003a, 2003b). Reporting patterns were explored in part to identify factors that may be used to predict reporting status.

CES survey distributions relative to the reporting pattern taxonomy developed earlier in this chapter are presented in Table 8. These results encompass the eighteen month period January 2001 through June 2002 and exclude Complete Nonresponse.

Table 8-Reporting Pattern Distributions

Reporting Pattern Distributions
Selected Industries, Jan '01 - Jun '02

	Manufacturing NAICS 31xx-33xx	Wholesale Trade NAICS 42xx-43xx	Mining NAICS 1133, 21xx	Construction NAICS 23xx
Complete Response	57.0%	49.8%	51.2%	47.4%
Strict Attrition	9.3%	11.8%	10.6%	9.4%
Strict Late Entry	7.0%	10.5%	10.2%	9.9%
Strict Late Entry Attrition	2.1%	3.5%	3.2%	3.5%
Attrition with Episodic Nonresponse	4.1%	4.6%	4.2%	5.4%
Late Entry with Episodic Nonresponse	2.0%	2.8%	2.0%	2.5%
Late Entry Attrition with Episodic Nonresponse	0.6%	0.6%	0.6%	0.4%
Strict Episodic Nonresponse	17.8%	16.5%	17.9%	21.5%

Roughly half the sample provided complete response for the eighteen month period. These units were thus able to be used in the 3rd closing estimates for all months. Other first order reporting patterns (Strict Attrition, Strict Late Entry, Strict Episodic Nonresponse) account for just over one-third of the sample.

Attrition (classified based on observing reporting patterns through December 2002) occurred for 15% - 20% of the sample, while some type of episodic nonresponse occurred for roughly 25% of the sample. (Note: Late entry could not be distinguished from initiation of new sample units (carried out on a flow basis); thus, some establishments classified as late entry may actually belong to the next higher level. In addition, some establishments classified as attrition may have become out of business.)

For complete response, as well as for attrition and late entry (during their period of reporting), timeliness of reporting affects which closing the sample units are used in.

For episodic reporting, any gaps result in the sample unit being unusable for the month of nonreporting as well as the first month of reporting following a gap.

1. Attrition

A second portion of nonresponse in a panel survey is due to sample establishments that stop reporting as of some point in time. Rosen, et al. (1993) classified attrition for the CES survey as: establishment went out of business; establishment overtly refused to continue participation; and establishment simply ceased reporting. Reasons for refusal and ceasing reporting include fatigue and, for establishment surveys, change in contact person within the establishment, with the result that a new decision is made relative to survey participation. CES guidelines treat reporting gaps of six months as attrition.

Data for attritors are not utilized in the weighted link relative estimator, with the implicit assumption being that the growth rate from month $t-1$ to t is the same for attritors as for available reporters within estimation cell. To the extent this assumption fails to hold, the accuracy of the CES survey estimates will be adversely affected.

A cumulative attrition rate through month t may be calculated as

$$Att\%_{1,t} = \frac{\sum_{t=1}^{t} n_{Att,t*}}{n_{Act,1}} \times 100\%$$

where

$n_{Att,t*}$ is the number of sample establishments becoming attritors effective month $t*$

$n_{Act,1}$ is the number of active sample establishments as of month 1

Cumulative attrition rates by major industry segment for the period January 2001 through June 2002 are presented in Figure 4, relative to active sample establishments as of December 2000. Attrition rates weighted by employment, are provided in Figure 5.

Figure 4-Cumulative Attrition Rate (unweighted)

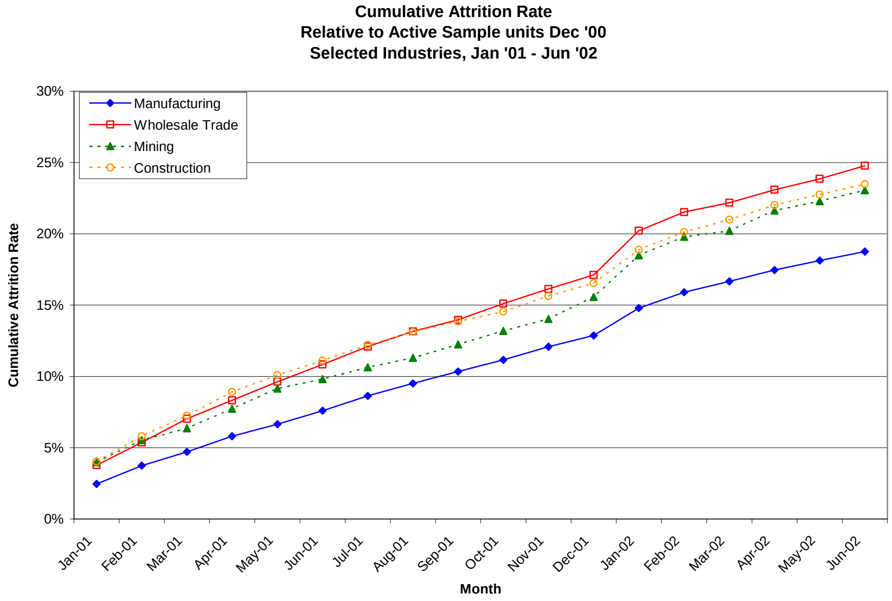
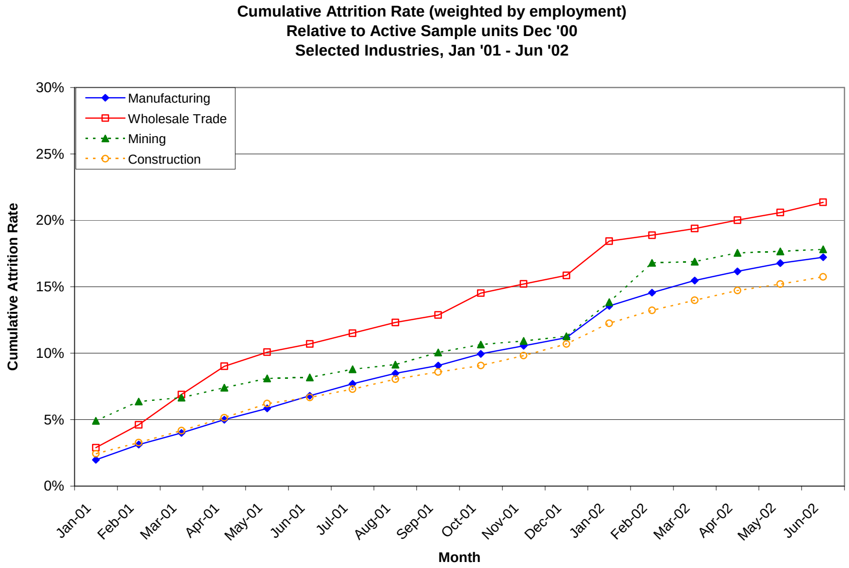


Figure 5-Cumulative Attrition Rate (weighted)



These graphs suggest cumulative attrition rates at the establishment level were slightly less for Manufacturing, while cumulative attrition rates weighted by employment tended to be slightly greater for Wholesale Trade. These data also provide an indication that Attritors tend to be smaller establishments, as the cumulative attrition rate is greater for establishments than for employment. Again, this is consistent with CES operational procedures which place greater emphasis on ensuring continued participation of larger establishments, so as to control the impact on survey estimates. This result may also be due in part to a greater likelihood of smaller establishments to go out of business, which could not be distinguished from attrition in this analysis.

A monthly attrition rate for month t may be calculated as

$$Att\%_T = \frac{n_{Att,T}}{n_{Act,1} - \sum_{t=1}^{T-1} n_{Att,t}} \times 100\%$$

Monthly attrition rates for the period January 2001 through June 2002, based on unweighted and weighted counts, respectively, are presented in Figure 6 and Figure 7. These graphs show attrition rates higher in January (2.2% - 4.0% for establishments and 1.7% - 4.9% for employment) than for the remaining months (0.5% - 1.9% for establishments and 0.1% - 3.4% for employment). Attrition rates are more variable for employment, especially for Mining.

The larger January attrition rate is likely due to the data collection process, in which establishments are mailed a calendar year log form in January. It is reasonable to assume some establishments opt to discontinue participation in the survey when they receive the new log form, as it provides a physical reminder of the expectations

BLS has for their continued participation in the survey for the next 12 months. There appears to be a potential carry-over of this attrition effect in February.

Figure 6-Monthly Attrition Rate (unweighted)

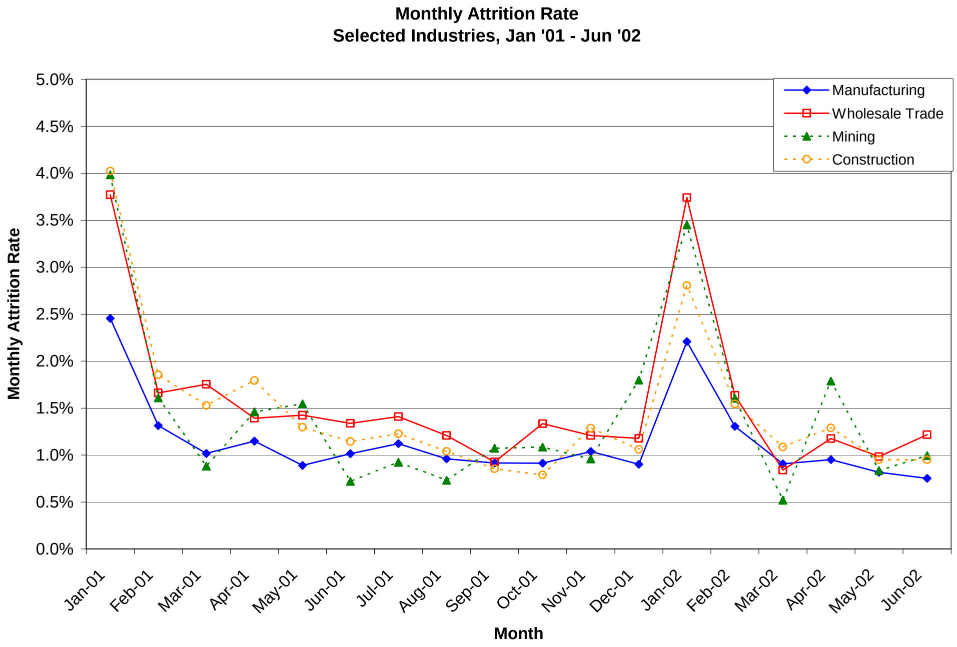
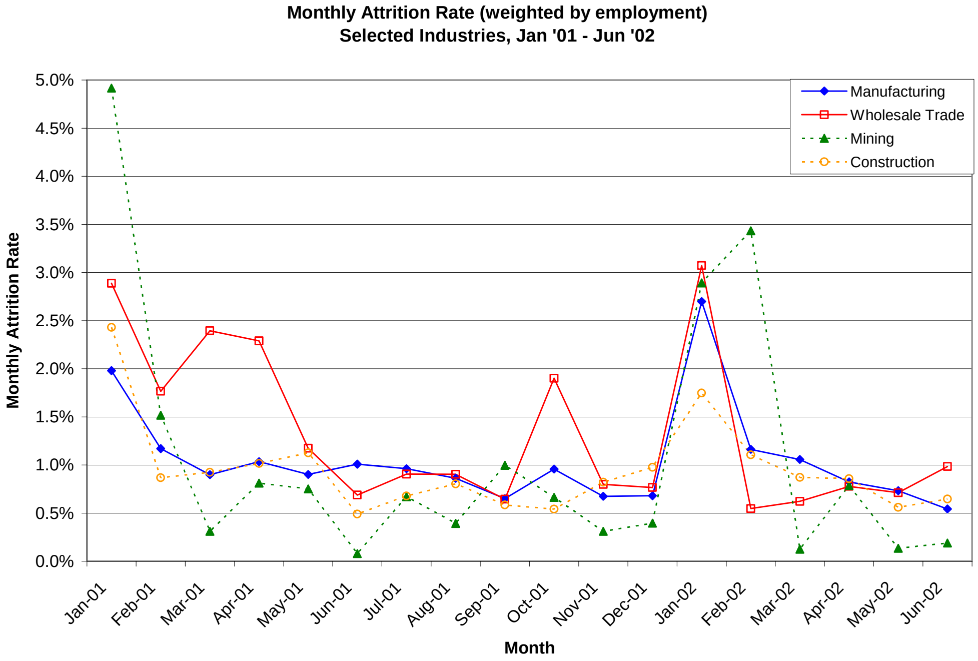


Figure 7-Monthly Attrition Rate (weighted)



2. Episodic Nonresponse

Episodic nonreporting represents sample establishments that do not report for a given month, but do report for a subsequent month. Gaps could be due to a variety of factors, such as change in data reporters, and seasonal closings. Episodic nonreporting can only be distinguished from attrition post hoc.

Episodic nonresponse may be viewed relative to the total sample size, with a within-month episodic nonresponse rate calculated as

$$ENR\%_t = \frac{n_{ENR,t}}{n_{Act,t}} \times 100\%$$

where

$n_{ENR,t}$ is the number of sample establishments that are episodic nonreporters in month t

$n_{Act,t}$ is the number of active sample establishments as of month t

Monthly episodic nonresponse rates for the period January 2001 – June 2002, based on unweighted and weighted counts, respectively, are presented in Figure 8 and Figure 9. These results show episodic nonresponse rates ranging from 1.2% to 5.1% for establishments and, excluding Mining, from 1.1% - 4.7% for employment. Mining episodic nonresponse rates for employment were much more variable, ranging from 0.6% to 9.0%. Thus, for episodic nonresponse rates, there do not appear to be any differences due to employment size.

Figure 8-Episodic Nonresponse Rate (unweighted)

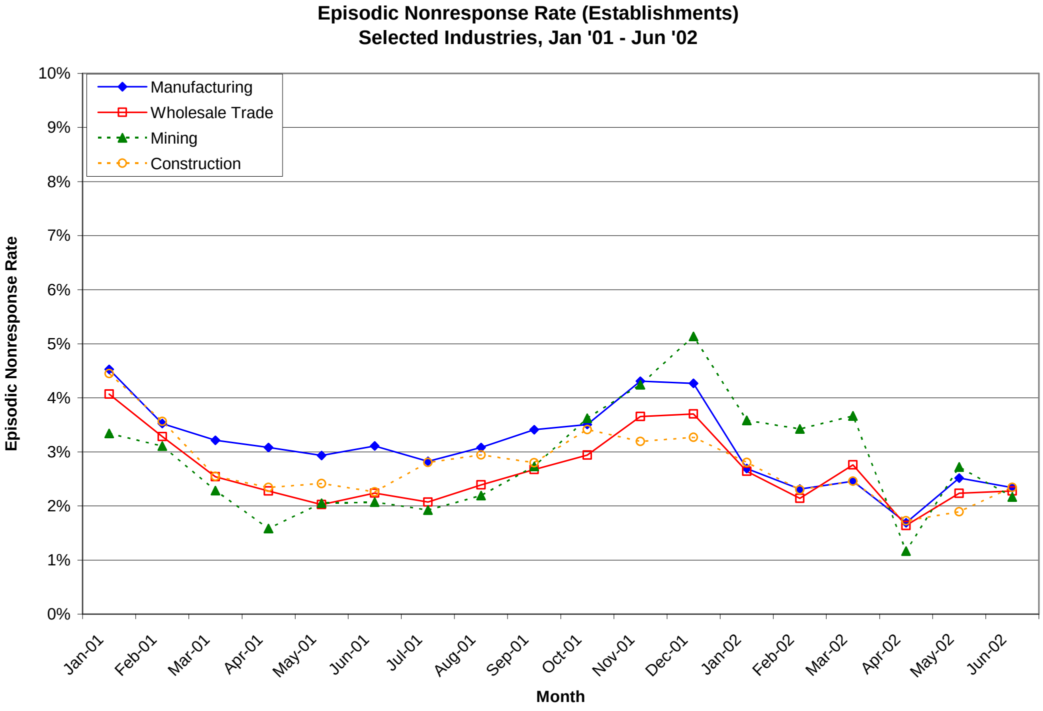
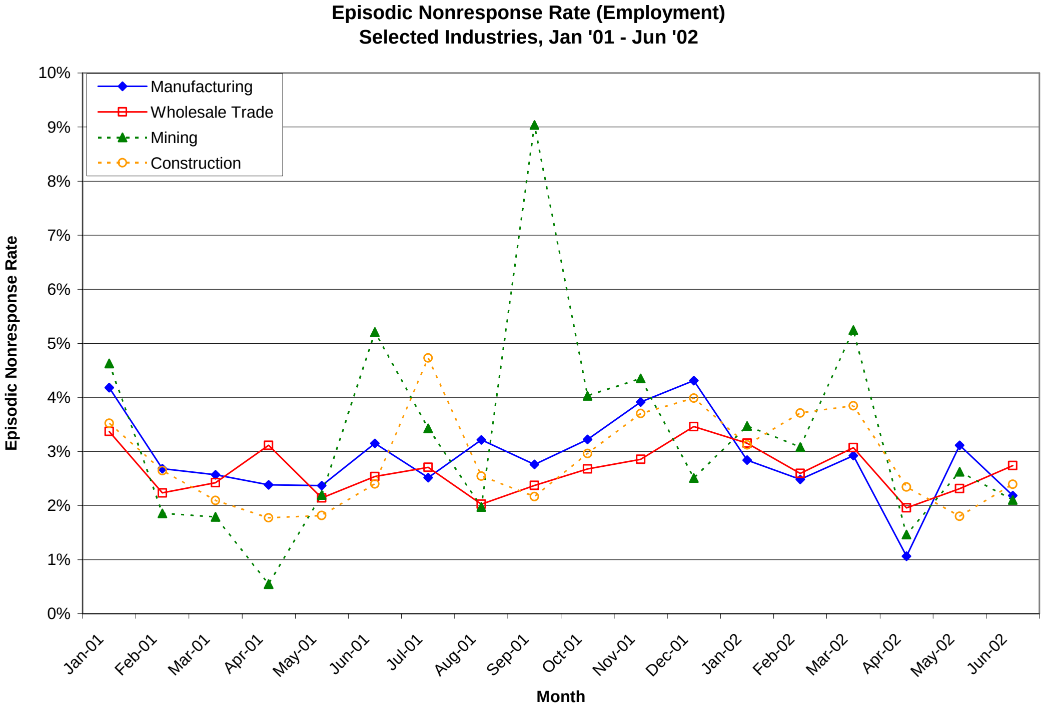


Figure 9-Episodic Nonresponse Rate (weighted)



The distribution of the maximum gap in nonreporting for episodic nonreporters in 2001 is presented in Table 9.

Table 9-Nonreporting Gaps

Nonreporting Gaps
Episodic Nonreporters in 2001

Longest gap for episodic nonreporters	Manufacturing	Wholesale Trade	Mining	Construction
1 month	42.7%	43.5%	49.3%	40.3%
2 months	21.2%	20.4%	19.7%	21.3%
3 months	17.3%	16.4%	13.5%	18.5%
4 months	11.0%	13.2%	8.8%	12.4%
5 months	5.9%	5.4%	6.2%	5.8%
6 months	1.9%	1.1%	2.6%	1.7%
7+ months	0.0%	0.0%	0.0%	0.0%

Between 40% and 49% of the episodic nonreporters experienced no more than a one month gap in nonreporting, while 18% - 20% experienced a gap of more than three months. Long gaps not leading to attrition may be a result of nonresponse conversion efforts undertaken for the CES survey.

Episodic nonreporting creates a carry-over effect in the use of a sample unit, due to the nature of the CES estimator. A sample establishment that does not report for a given month will be left out of the calculation of the weighted link relative not only for that month, but also for the succeeding month, as it will not be contained within the set of constant reporters.

3. Combined Nonreporting

The prior information about the components of nonresponse can be viewed as a whole across time. Such a picture can provide some insight into the nature of the problems faced in appropriately compensating for nonresponse.

Information about the distribution of the reporting behavior in 2001 for the active sample as of December 2000 is provided in Table 10.

Table 10-2001 Reporting Behavior

Reporting Behavior 2001
Active Sample Units as of Dec '00

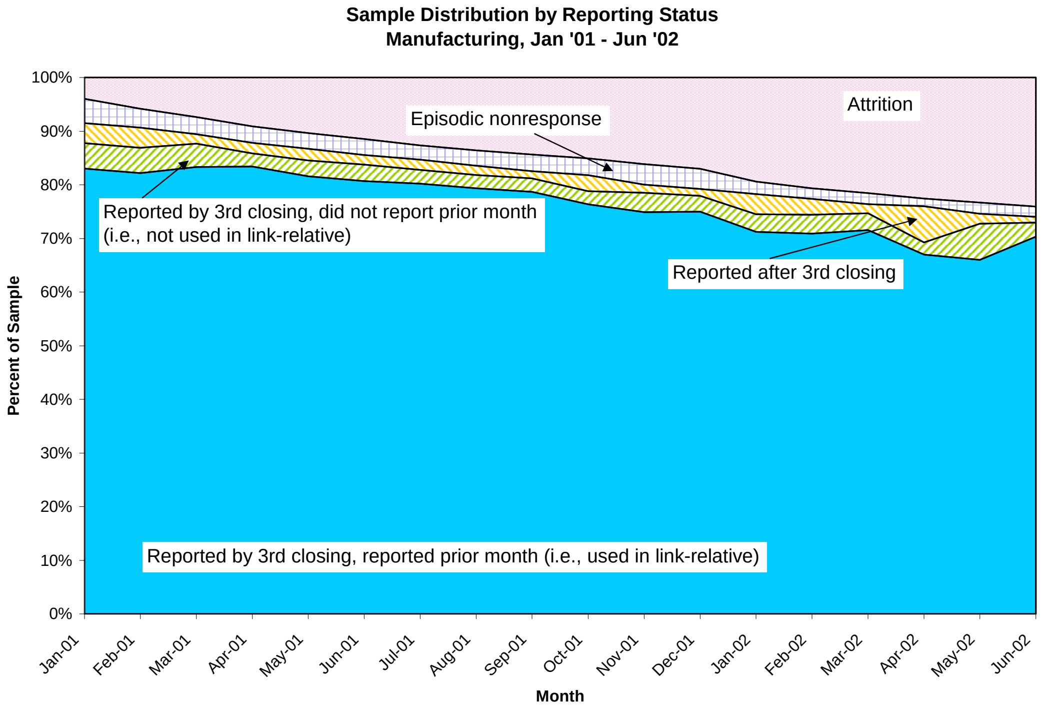
	Manufacturing	Wholesale Trade	Mining	Construction
Respond all 12 months	74.5%	69.4%	70.3%	68.7%
Attritor during 12 months	11.0%	15.0%	13.6%	13.0%
Episodic NR during 12 months	14.5%	15.6%	16.1%	18.3%

Roughly 70% of sample establishments reported all 12 months, while between 10% and 15% became attritors from the sample. The remaining 15% to 20% of sample units experienced at least one occasion of episodic nonresponse in the year.

Although roughly 15% of the sample had an episodic nonresponse occurrence in 2001, the frequency within a given month is somewhat less. The distribution of reporting status for Manufacturing from January 2001 through June 2002 is provided in Figure 10.

This graph shows episodic nonreporting accounted for less than 5% of the sample within a month. However, as stated earlier, episodic nonreporting also affects the usability of a subsequent month reporter, due to the need for two consecutive months of data for the weighted link relative. As seen from the diagonally hatched portion of the bar, this carry-over effect resulted in an additional 2% - 7% of the sample being unusable for the weighted link relative within a month. In addition, there are a small percentage of the sample establishments (1% - 7%) that report too late for inclusion even in the third closing estimates.

Figure 10-Sample Distribution by Reporting Status



B. Late Reporting

The current CES estimator only utilizes sample units reporting for both months t and $t-1$. For preliminary estimates, a sample unit must have reported by first closing for month t as well as have reported for month $t-1$. The sample is expanded for revised estimates with the inclusion of late reporters for month t that had reported for month $t-1$. Thus both preliminary (preliminary reporter vs. not preliminary reporter) and final reporting status (late reporter vs. nonreporter) impact on the use of sample unit in estimation.

For purposes of discussion, reporting status for reference period t for unit i may be summarized by

$$\mathbf{X}_{ti} = \left(X_{ti}^{PR} \quad X_{ti}^{LR} \quad X_{ti}^{NR} \right)^T$$

where the superscripts refer to preliminary reporting (PR), late reporting (LR), and nonresponse (NR)

$$X_{ti}^{PR} = \begin{cases} 1 \text{ if unit } i \text{ is a preliminary reporter for month } t \\ 0 \text{ otherwise} \end{cases}$$

$$X_{ti}^{LR} = \begin{cases} 1 \text{ if unit } i \text{ is a late reporter for month } t \\ 0 \text{ otherwise} \end{cases}$$

$$X_{ti}^{NR} = \begin{cases} 1 \text{ if unit } i \text{ is a nonreporter for month } t \\ 0 \text{ otherwise} \end{cases}$$

Frequency of occurrence for reporting patterns yielding at least one month of reported data is provided in Table 11. As this table shows, for preliminary estimates the current CES estimator is only able to utilize data for roughly three-fourths of the sample units for which data are available for at least one of the two months. Sample units for which only prior month's data are available account for roughly 90% of the remaining sample units. Sample units eventually classified as late reporters account for roughly 75% of the subset for which only prior month's data are available.

Table 11-Reporting Pattern Distribution

Reporting Patterns
Frequency of Occurrence Jan 2001 - June 2002

Reporting Pattern		Manufacturing	Wholesale Trade	Mining	Construction
Month t	Month t-1				
$X_{ti}^{PR}=1$	$X_{(t-1)i}^{NR}=0$	76.0%	69.3%	71.9%	78.7%
$X_{ti}^{PR}=1$	$X_{(t-1)i}^{NR}=1$	2.4%	2.5%	2.6%	3.1%
$X_{ti}^{PR}=0$	$X_{(t-1)i}^{NR}=0$	21.6%	28.2%	25.4%	18.2%
$X_{ti}^{LR}=1$	$X_{(t-1)i}^{NR}=0$	16.5%	23.1%	19.5%	12.3%
$X_{ti}^{NR}=1$	$X_{(t-1)i}^{NR}=0$	5.2%	5.1%	5.9%	5.9%

1. Timeliness of Reporting Across Time

Timeliness of reporting is an issue for most sample establishments in the CES survey, although not on a continual basis. A top-level distribution of frequency of first-closing reporting for establishments in the Complete Response reporting pattern, for the eighteen-month period January 2001 – June 2002, is presented in Table 12.

The proportion of establishments in the Complete Response reporting pattern that reported on-time every month ranged from 23% to 29% at the industry level, while the proportion of establishments that reported late every month ranged from 1% to 12%. Thus, the majority of sample establishments vary in terms of which closing their data are used in.

Table 12-Timeliness of Reporting Pattern Distributions

Timeliness of Reporting Pattern Distributions
Selected Industries, Jan '01 - Jun '02
Sample Reporting all Eighteen Months

	Manufacturing NAICS 31xx-33xx	Wholesale Trade NAICS 42xx-43xx	Mining NAICS 1133, 21xx	Construction NAICS 23xx
Every Month by First Closing	27.7%	22.7%	22.8%	29.1%
12 - 17 Months by First Closing	55.4%	53.6%	51.3%	60.3%
6 - 11 Months by First Closng	10.6%	8.3%	16.2%	8.1%
1 - 5 Months by First Closing	4.0%	3.1%	7.6%	2.0%
No Month by First Closing	2.2%	12.3%	2.1%	0.6%

2. Late Reporting vs. Preliminary Reporting

This set of tables looks at late reporting rates as a proportion of total reporting, by selected characteristics. These were characteristics previously mentioned as related to late reporting (number of reporting days, size, length of pay period), as well as other factors potentially related to late reporting (prior reporting behavior).

Prior reporting behavior may be indicative of current behavior for sample units. With information on reporting status available across time, it is possible to examine relationships between reporting status in recent months and reporting status for the current month. In particular, late reporting in a recent month was hypothesized to be correlated with late reporting in the current month.

For a variety of reasons, some sample establishments are unable to respond within the narrow timeframe required for publication of first closing results, but do provide data for the survey month at a later point in time (Rosen, et al., 1991). Calendar effects appear to play a role in late reporting. For the CES survey, the number of reporting days available for data collection depends upon the day of the week the 12th of the month falls on; the shorter the data collection period, the greater the likelihood for late reporting. In addition, as data are to be reported for the pay period containing the 12th day of the month, the length of a sample establishment's pay period could affect availability of the information to be reported

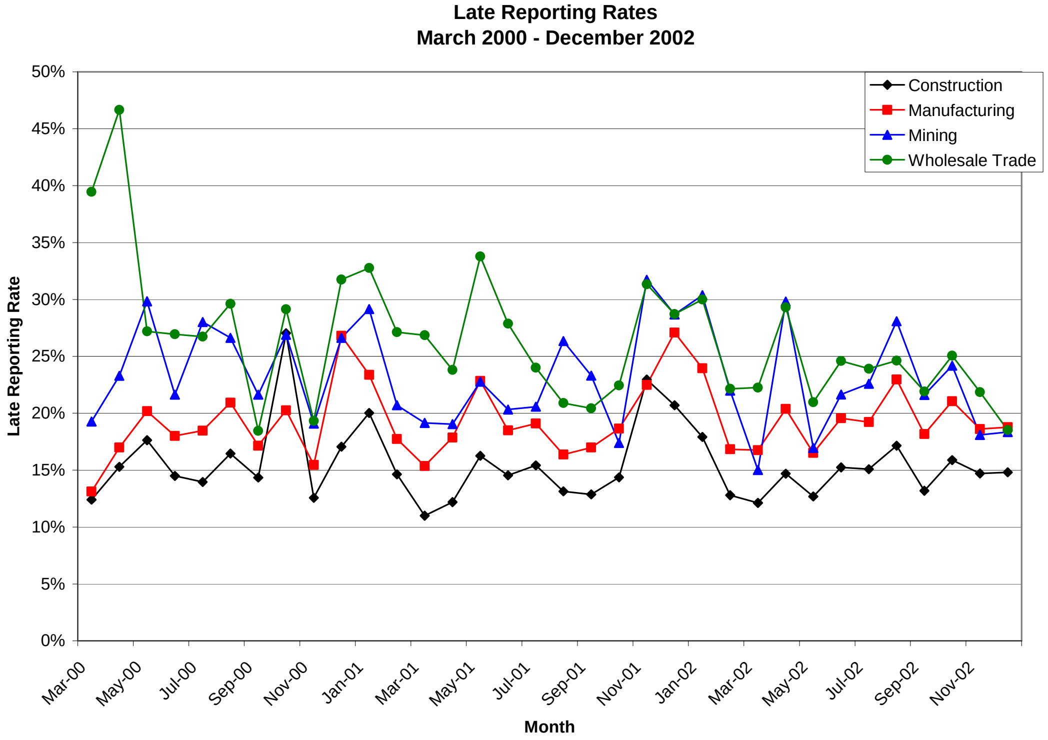
While the data for these late reporters are utilized in second and third closing estimates (depending upon when they report), any differences between their month-to-month trends and that assumed by the weighted link relative estimator will drive the direction and magnitude of revisions to the first closing estimates.

A late reporting rate, conditional on reporting, may be calculated as

$$LR\% \mid \left(X_{ti}^{NR} = 0\right) = \frac{\sum_i X_{ti}^{LR}}{\sum_i \left(X_{ti}^{LR} + X_{ti}^{PR}\right)} \times 100\%$$

Late reporting rate, conditional on reporting, for the period March 2000 through December 2002 are presented in Figure 11. These graphs show late reporting rates have generally ranged between 10% and 35%. This percentage varies across both time and industry.

Figure 11-CES Late Reporting Rates



Late reporting rates were then examined by various factors felt to be related to timeliness of reporting – design size class, length of pay period, number of reporting days, prior two months' reporting status, and calendar month. Results are provided in Table 13.

The results suggest late reporting rates are greater for larger establishments, establishments with a monthly pay period, and establishments which had been either a late reporter or nonrespondent the prior months. To a lesser degree, months with fewer reporting days exhibit higher late reporting rates, as does the month of January.

Table 13-Late Reporting Rates, Conditional on Reporting, for Selected Characteristics

Late Reporting Rates, Conditional on Reporting
by Design Size Class
(3/00 - 12/02)

Design Size Class	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
<10	14.32%	3.07%	13.12%	2.35%	17.44%	7.27%	18.53%	14.33%
10-19	14.26%	3.99%	12.62%	2.48%	16.75%	4.29%	17.34%	10.87%
20-49	13.89%	4.21%	13.88%	2.43%	19.09%	5.15%	19.29%	8.31%
50-99	15.20%	5.46%	16.49%	3.40%	20.80%	5.49%	20.36%	6.55%
100-249	15.32%	3.51%	17.38%	3.08%	26.79%	6.18%	24.01%	4.76%
250-499	19.56%	3.60%	19.24%	3.45%	29.10%	12.69%	27.88%	4.92%
500-999	21.16%	5.37%	23.41%	4.72%	22.90%	8.99%	35.42%	6.99%
1000+	32.97%	6.78%	27.75%	4.43%	32.00%	11.33%	48.30%	9.22%

Late Reporting Rates, Conditional on Reporting
by Length of Pay Period
(3/00 - 12/02)

Length of Pay Period	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
Weekly	14.15%	3.30%	16.84%	2.80%	19.99%	5.56%	24.89%	6.01%
Bi-Weekly	18.29%	4.05%	23.86%	4.73%	26.27%	4.81%	20.31%	8.64%
Semi-Monthly	19.00%	5.26%	20.56%	4.77%	22.77%	9.44%	22.75%	6.95%
Monthly	39.06%	5.78%	44.78%	5.05%	41.22%	8.07%	58.17%	6.44%

Late Reporting Rates, Conditional on Reporting
by Number of Reporting Days
(3/00 - 12/02)

Number of Reporting Days	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
9	16.94%	0.98%	21.51%	1.88%	26.30%	4.98%	30.49%	4.65%
10	16.89%	2.42%	21.83%	2.81%	26.10%	4.48%	27.73%	3.53%
11	16.68%	5.36%	20.29%	3.31%	22.73%	3.26%	25.87%	3.75%
12	14.28%	0.92%	17.91%	0.85%	22.60%	3.74%	28.71%	10.52%
13	14.18%	1.94%	18.34%	2.63%	21.67%	4.63%	22.89%	3.17%
14	14.49%	4.31%	17.32%	2.67%	22.32%	5.60%	24.42%	4.72%
15	13.61%	1.70%	15.94%	4.00%	18.80%	0.66%	28.99%	14.80%

Late Reporting Rates, Conditional on Reporting
by Prior 2 Months' Reporting Pattern
(3/00 - 12/02)

Prior 2 Months' Reporting Pattern		Construction		Manufacturing		Mining		Wholesale Trade	
Month t-1	Month t-2	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
PR	PR	7.43%	2.57%	8.18%	2.58%	9.12%	4.06%	8.63%	4.23%
	LR	25.32%	5.35%	30.06%	6.66%	29.32%	11.92%	28.33%	11.45%
	NR	16.04%	4.43%	17.36%	4.95%	18.12%	15.41%	16.91%	10.89%
LR	PR	26.75%	5.71%	31.25%	7.03%	40.49%	16.14%	30.58%	13.76%
	LR	56.10%	7.39%	66.45%	7.71%	67.37%	14.41%	81.33%	12.51%
	NR	34.36%	7.19%	42.05%	7.74%	45.40%	15.92%	47.24%	14.60%
NR	PR	40.88%	10.79%	46.80%	9.73%	45.04%	17.20%	41.13%	12.52%
	LR	66.72%	8.47%	76.96%	6.86%	74.24%	18.30%	71.20%	13.03%
	NR	61.24%	12.18%	67.25%	7.15%	66.05%	18.26%	59.21%	12.45%

Late Reporting Rates, Conditional on Reporting
by Calendar Month
(3/00 - 12/02)

Month	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
Jan	18.97%	1.49%	23.66%	0.41%	29.77%	0.85%	31.38%	1.96%
Feb	13.71%	1.30%	17.28%	0.64%	21.34%	0.92%	24.64%	3.53%
Mar	11.84%	0.75%	15.08%	1.84%	17.81%	2.43%	29.52%	8.90%
Apr	14.06%	1.65%	18.41%	1.76%	24.05%	5.43%	33.26%	11.93%
May	15.53%	2.55%	19.85%	3.18%	23.18%	6.45%	27.31%	6.41%
Jun	14.75%	0.42%	18.68%	0.80%	21.20%	0.76%	26.48%	1.69%
Jul	14.82%	0.76%	18.93%	0.40%	23.73%	3.85%	24.88%	1.60%
Aug	15.58%	2.14%	20.09%	3.39%	27.02%	0.93%	25.05%	4.37%
Sep	13.47%	0.78%	17.44%	0.64%	22.18%	0.97%	20.27%	1.74%
Oct	19.10%	6.92%	19.98%	1.23%	22.82%	4.90%	25.55%	3.37%
Nov	16.74%	5.49%	18.85%	3.52%	22.97%	7.59%	24.17%	6.33%
Dec	17.52%	2.97%	24.22%	4.72%	24.55%	5.47%	26.33%	6.93%

The result for design size is consistent with operational procedures used in CES data collection, wherein more emphasis is placed on obtaining responses from larger establishments, and also with operational aspects of reporting, wherein large establishments reporting for multiple worksites may find it difficult to compile all the information in time for first closing. Likewise the result for length of pay period is consistent with operational aspects of reporting, as establishments with monthly pay periods generally would not have data for the reference pay period until late in or after the close of the collection period.

Relationship between prior late reporting and increased late reporting rates likely indicates ability of a sample establishment to obtain the required information within the collection period. This factor may also be correlated with length of pay period.

One reason for the relatively weak relationships with late reporting rates for number of reporting days and calendar maybe potentially more complex relationships involving calendar dynamics. Rather than the number of reporting days, it may be that closing date in conjunction with length of pay period may affect late reporting rates. For example, the likelihood of late reporting for an establishment with a bi-weekly pay period could be greater is the week containing the 12th of the month were the first week of the pay period than if it were the second week of the pay period. These types of relationships were not investigated as part of this research.

These findings suggest inclusion of establishment design size, length of pay period and recent reporting status as factors in predicting late reporting rates at the establishment level.

3. Late Reporting vs. Nonresponse

For purposes of developing a model to predict final reporting status, the remaining tables look at conditional late reporting rates relative to the same factors as in the previous section. Conditional late reporting rate, given a sample unit was not a preliminary reporter, was defined as

$$LR\% \mid \left(X_{ti}^{PR} = 0\right) = \frac{\sum_i X_{ti}^{LR}}{\sum_i \left(X_{ti}^{LR} + X_{ti}^{NR}\right)} \times 100\%$$

Length of time from last report can be expected to be strongly correlated with likelihood of reporting in the current period. As evidenced in Table 14 this is true for the CES sample. Sample units with a gap in reporting of four or more months averaged less than a 10% conditional late reporting rate for each industry. Reporting gap was felt to be such a dominant factor that the profile relative to other factors was carried out conditional on a top-level classification of reporting gap of 3 months or less.

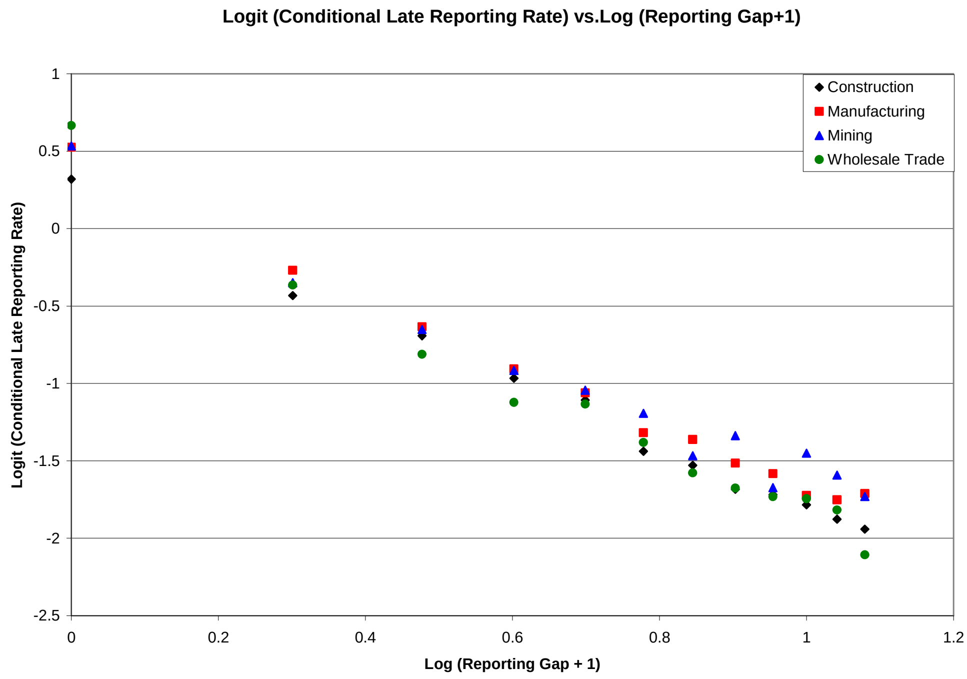
Table 14-Conditional LR Rates, by Reporting Gap

Late Reporting Rates, Conditional on Not Preliminary Reporter
by Nonreporting Gap
(3/00 - 12/02)

Nonreporting Gap (in Months)	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
0	67.61%	7.20%	77.03%	4.55%	77.35%	6.76%	82.27%	4.15%
1	26.98%	5.68%	34.97%	4.90%	30.95%	11.01%	30.14%	11.42%
2	16.88%	10.70%	18.84%	3.54%	18.25%	13.36%	13.38%	6.12%
3	9.76%	7.61%	11.05%	3.32%	10.82%	13.50%	7.02%	3.60%
4	7.25%	9.94%	8.01%	5.00%	8.29%	11.05%	6.85%	6.25%
5	3.52%	2.60%	4.59%	2.22%	6.03%	9.53%	4.00%	5.13%
6	2.88%	1.62%	4.16%	2.05%	3.30%	4.92%	2.58%	2.15%
7	2.03%	1.25%	2.97%	2.07%	4.40%	9.44%	2.07%	1.75%
8	1.88%	1.56%	2.55%	1.85%	2.08%	3.75%	1.82%	2.43%
9	1.62%	1.26%	1.86%	1.50%	3.42%	5.49%	1.77%	2.23%
10	1.31%	0.99%	1.74%	1.47%	2.50%	4.43%	1.50%	2.98%
11	1.13%	1.27%	1.91%	2.05%	1.83%	5.32%	0.78%	0.83%
12+	0.96%	0.60%	1.03%	0.66%	0.38%	0.42%	0.59%	0.41%

By examining the nature of the relationship between conditional late reporting rates and reporting gaps, it appears a transformation to the logit of the conditional late reporting rate and the log of one plus the length of the reporting gap follow a linear relationship, as evidenced by Figure 12.

Figure 12-Logit (Conditional LR Rate) vs Log(Gap+1)



As a logistic regression would be a reasonable model for the conditional late reporting rate, as discussed more fully in section D, these results suggest inclusion of the log transformation of the length of reporting gap in the model for predicting current month reporting status.

Prior reporting behavior for a sample unit was hypothesized to be related to conditional late reporting rate. In particular, late reporting in a recent month was hypothesized to be correlated with late reporting in the current month. Information on the conditional late reporting rate relative to reporting status for the prior two months, excluding sample units with a reporting gap of 4+ months, is provided in Table 15. As can be seen, higher conditional late reporting rates are associated with prior reporting (both preliminary and late), with prior late reporting associated with higher conditional late reporting rates than prior preliminary reporting, especially when the late reporting occurred in month $t-1$. These results suggest inclusion of prior reporting patterns in the model for predicting current month reporting status.

Table 15-Conditional LR Rates, by Prior Reporting Pattern

Late Reporting Rates, Conditional on Not Preliminary Reporter
by Prior 2 Months' Reporting Pattern
Excluding Sample with Reporting Gap 4+ Months
(3/00 - 12/02)

Prior 2 Months' Reporting Pattern		Construction		Manufacturing		Mining		Wholesale Trade	
Month t-1	Month t-2	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
PR	PR	63.95%	10.96%	72.84%	7.94%	69.25%	12.63%	69.62%	9.72%
	LR	65.80%	6.38%	73.57%	4.98%	72.53%	13.06%	73.73%	7.24%
	NR	40.43%	8.96%	43.04%	7.69%	44.58%	28.91%	42.17%	16.75%
LR	PR	77.50%	4.92%	83.17%	4.18%	83.99%	10.88%	82.95%	7.38%
	LR	80.71%	4.14%	87.05%	2.43%	86.87%	7.65%	92.77%	2.62%
	NR	61.98%	7.04%	65.45%	6.14%	69.86%	13.07%	69.71%	11.60%
NR	PR	23.96%	6.11%	30.32%	5.77%	25.91%	10.96%	25.51%	12.50%
	LR	34.38%	6.73%	41.87%	5.18%	40.04%	18.33%	37.98%	12.35%
	NR	13.76%	9.55%	15.50%	3.21%	15.33%	12.43%	10.65%	5.27%

Two characteristics of establishments were hypothesized to be related to conditional reporting rates, length of payroll and prior months' employment trend.

Payroll structure affects late reporting, given the nature of the reference period (pay period containing the 12th of the month) and the reporting period (which closes on the Friday two weeks after the end of the week containing the 12th). Sample units with monthly pay periods will likely not have data available within the reporting period. Sample units with bi-weekly pay periods will be faced with varying abilities to have data for reporting, depending upon when their pay period ends relative to the 12th. Sample units with weekly and semi-monthly pay periods could be expected to be most likely to be able to report within the prescribed reporting period. This supposition is relatively supported by conditional late reporting rates by length of pay period, excluding sample units with a reporting gap of 4+ months (Table 16). These results suggest inclusion of length of payroll in the model for predicting current month reporting status.

Table 16-Conditional LR Rates, by LOPP

Late Reporting Rates, Conditional on Not Preliminary Reporter
by Length of Pay Period
Excluding Sample with Reporting Gap 4+ Months
(3/00 - 12/02)

Length of Pay Period	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
Weekly	50.24%	8.96%	61.41%	4.93%	60.74%	10.95%	64.77%	7.85%
Bi-Weekly	53.23%	7.56%	66.70%	4.95%	63.10%	8.36%	63.00%	10.97%
Semi-Monthly	43.57%	8.36%	59.35%	7.76%	60.53%	12.63%	60.43%	8.82%
Monthly	64.49%	6.31%	76.06%	4.19%	68.57%	7.06%	83.33%	4.10%

The prior month’s employment trend was hypothesized to be related to conditional late reporting, in that respondents in sample units experiencing large declines may be more focused on business issues than reporting data for a survey, and thus may have lower conditional late reporting rates. Sample units were rank ordered based on prior month’s employment trend (change for establishments with <50 reported employment, to avoid unstable growth rates, and growth rate for establishments with

50+ reported employment, to avoid unstable change). As shown in Table 17, no evidence of an effect due to prior month employment trend was seen. The much lower conditional late reporting rates seen for establishments with an unknown ranking was felt to be related to nonreporting.

Table 17-Conditional LR Rates, by Prior Employment Change

Late Reporting Rates, Conditional on Not Preliminary Reporter
by Prior Month's Employment Change
Excluding Sample with Reporting Gap 4+ Months
(3/00 - 12/02)

Prior Month Employment <50

Ranked Prior Month's Employment Change	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
N/A	12.54%	3.43%	13.22%	3.67%	13.79%	7.45%	14.19%	8.96%
Bottom Third	65.63%	11.73%	72.23%	4.93%	69.31%	11.37%	81.27%	5.72%
Middle Third	62.63%	9.65%	72.53%	5.31%	70.95%	10.19%	81.90%	5.82%
Top Third	66.67%	12.55%	74.36%	5.59%	74.01%	9.19%	82.01%	5.38%

Prior Month Employment 50+

Ranked Prior Month's Employment Growth Rate	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
N/A	52.30%	10.71%	54.49%	8.20%	57.51%	29.29%	55.28%	14.95%
Bottom Third	73.09%	8.89%	77.91%	4.83%	81.38%	7.91%	83.47%	4.34%
Middle Third	72.77%	8.90%	78.78%	4.86%	78.78%	7.92%	84.28%	4.45%
Top Third	72.74%	9.30%	78.43%	4.85%	81.18%	6.54%	83.28%	4.25%

Both late reporting and nonreporting are affected by operational aspects of the CES survey. Given the importance of larger units, more emphasis is placed upon achieving high preliminary reporting rates as well as high overall reporting rates for larger units. Results in Table 18 indicate average conditional late reporting rates by design size class, excluding sample units with a reporting gap of 4+ months, increase as establishment size increases, reflecting the relative effort placed on data collection by establishment size, as well as the greater likelihood of smaller establishments going out of business.

Table 18-Conditional LR Rates, by Design Size Class

Late Reporting Rates, Conditional on Not Preliminary Reporter
by Design Size Class
Excluding Sample with Reporting Gap 4+ Months
(3/00 - 12/02)

Design Size Class	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
<10	39.64%	7.48%	45.76%	6.89%	45.57%	16.53%	48.98%	11.22%
10-19	44.65%	8.86%	47.54%	6.86%	57.26%	12.04%	53.47%	9.98%
20-49	48.96%	10.40%	52.26%	6.26%	58.16%	9.07%	58.94%	9.29%
50-99	57.30%	9.35%	58.83%	5.37%	62.47%	10.31%	62.65%	8.58%
100-249	59.50%	7.54%	64.45%	5.82%	70.34%	9.35%	69.39%	6.35%
250-499	60.82%	6.17%	64.53%	5.53%	61.28%	18.77%	72.51%	6.25%
500-999	57.24%	9.60%	66.96%	6.18%	63.13%	19.62%	75.11%	8.56%
1000+	68.49%	8.82%	70.51%	4.24%	70.98%	14.11%	81.51%	3.94%

Average conditional late reporting rates by size class based on prior month reported employment, excluding sample units with a reporting gap of 4+ months, were also examined, and are provided in Table 19. While these also follow a roughly increasing function as establishment size increases, differences are less pronounced than for design size class. This is a logical outcome, as design size class is the operational information most readily available for which to prioritize nonreporting followup. The results on conditional late reporting rates suggest some measure of establishment size in the model for predicting current month reporting status.

Table 19-Conditional LR Rates, by Prior Month Size Class

Late Reporting Rates, Conditional on Not Preliminary Reporter
by Prior Month Reported Employment Size Class
Excluding Sample with Reporting Gap 4+ Months
(3/00 - 12/02)

Prior Month Reported Employment	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
NR	8.69%	2.61%	11.11%	3.60%	8.33%	5.22%	9.23%	5.51%
<10	58.86%	9.15%	67.78%	5.72%	66.21%	10.75%	78.41%	6.91%
10-19	64.24%	11.85%	72.19%	6.23%	73.90%	10.50%	81.29%	5.43%
20-49	68.51%	12.50%	73.12%	5.14%	72.13%	8.59%	82.54%	5.00%
50-99	71.11%	10.21%	75.83%	5.17%	77.15%	8.17%	82.61%	4.36%
100-249	72.18%	8.39%	77.96%	5.06%	81.13%	7.60%	83.01%	4.18%
250-499	73.24%	8.05%	76.33%	6.15%	78.54%	8.48%	82.85%	5.18%
500-999	71.88%	13.56%	77.55%	4.79%	80.56%	16.64%	80.37%	7.99%
1000+	74.03%	12.74%	80.38%	5.23%	85.07%	10.89%	78.60%	9.26%

Finally, conditional late reporting rates were examined relative to calendar effects, or what might be termed environmental factors.

The length of the data reporting period for a month depends upon the day of the week on which the 12th of the month falls. Data reporting periods vary from 9 to 15 days. One could expect months with shorter reporting periods to experience higher conditional late reporting rates. Calendar month was also examined to determine whether any evidence existed to support some type of effect on conditional late reporting rates. One could expect higher conditional late reporting rates associated with December, as respondents may be out of the office during much of the reporting period.

Interestingly, while results by number of reporting days show some evidence of higher conditional late reporting rates for months with 9 reporting days, months with 15 reporting days likewise showed some evidence of higher conditional late reporting rates, as shown in Table 20. This may be due to an interaction with a calendar month effect, as only two months had 15 reporting days, one of which was December, 2002.

Table 20-Conditional LR Rates, by Number of Reporting Days

Late Reporting Rates, Conditional on Not Preliminary Reporter
by Number of Reporting Days
Excluding Sample with Reporting Gap 4+ Months
(3/00 - 12/02)

Number of Reporting Days	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
9	53.20%	1.12%	67.75%	0.73%	68.81%	8.88%	72.07%	3.57%
10	50.41%	9.38%	62.72%	6.10%	60.17%	5.40%	66.73%	4.15%
11	55.67%	7.23%	65.82%	2.46%	64.80%	5.59%	68.56%	2.87%
12	50.78%	7.11%	64.60%	2.91%	64.66%	8.88%	70.67%	7.91%
13	44.66%	10.38%	62.16%	6.08%	57.67%	8.89%	63.79%	5.70%
14	50.19%	7.00%	60.00%	2.29%	61.00%	10.16%	65.80%	4.08%
15	54.41%	2.80%	64.09%	3.82%	69.08%	9.16%	74.66%	16.16%

Conditional late reporting dates by month, however, did not suggest a higher conditional late reporting rate for December nor for any other month, as seen in Table 21.

Table 21-Conditional LR Rates, by Calendar Month

Late Reporting Rates, Conditional on Not Preliminary Reporter
by Calendar Month
Excluding Sample with Reporting Gap 4+ Months
(3/00 - 12/02)

Month	Construction		Manufacturing		Mining		Wholesale Trade	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
Jan	51.57%	1.83%	63.32%	0.83%	60.10%	2.97%	65.07%	3.71%
Feb	44.42%	0.62%	58.13%	1.16%	51.97%	2.41%	62.79%	4.28%
Mar	46.72%	5.19%	60.82%	5.31%	60.13%	15.09%	71.26%	13.15%
Apr	45.24%	6.21%	60.08%	9.06%	63.62%	8.26%	71.30%	11.11%
May	51.93%	2.34%	65.31%	4.26%	63.45%	11.21%	69.11%	5.72%
Jun	50.99%	8.95%	65.03%	1.68%	64.52%	3.94%	68.99%	2.50%
Jul	50.53%	10.55%	65.23%	0.59%	64.18%	1.07%	66.63%	1.79%
Aug	49.49%	13.74%	66.51%	4.72%	66.71%	8.36%	69.02%	3.86%
Sep	47.68%	13.68%	63.95%	2.82%	63.72%	7.47%	65.10%	5.77%
Oct	56.96%	8.38%	63.65%	5.30%	61.34%	11.14%	67.26%	7.15%
Nov	55.83%	6.57%	59.93%	2.98%	60.42%	11.62%	67.42%	6.02%
Dec	57.58%	4.31%	66.59%	4.56%	65.07%	4.11%	67.27%	3.76%

4. Summary

The data on CES reporting patterns show late reporting to constitute a relatively large proportion of total reporting. Factors related to late reporting appears related to prior months' reporting status, length of pay period, and design size class.

D. Model for Predicting Final Reporting Status

An ancillary objective of this dissertation research was to specify a model for predicting reporting status for month t at the unit level, given preliminary reporting status is known (i.e., as of d_t, the cutoff date for month t). This information could be used to predict final reporting rates for month t (and thereby provide early warnings), identify areas of focus for followup efforts, and possibly allow early assessment of potential differences between preliminary and final estimates for month t. In addition, the reporting status model could potentially be integrated with the employment growth model to provide imputation representing (expected) predicted late reporters specifically.

At the time preliminary estimates are generated, reporting status for month t is not fully known. Reporting status is known for preliminary reporters, i.e., $\mathbf{X}_{tci} = \begin{pmatrix}1 & 0 & 0\end{pmatrix}^T$, but for the remainder of the sample units it is unknown, with two possible outcomes, late reporter or nonresponse, i.e., $\mathbf{X}_{tci} = \begin{pmatrix}0 & 1 & 0\end{pmatrix}^T, \begin{pmatrix}0 & 0 & 1\end{pmatrix}^T$. The model seeks to predict month t reporting status for sample units with unknown reporting status as of preliminary cutoff date for month t, i.e., those units for which

$$\mathbf{X}_{tci} = \begin{pmatrix}0 \\ . \\ .\end{pmatrix}$$

For a sample unit, there are three states that can occur relative to reporting status – preliminary reporting $\left(X_{tci}^{PR} = 1\right)$, late reporting $\left(X_{tci}^{LR} = 1\right)$, or nonresponse $\left(X_{tci}^{NR} = 1\right)$. The vector of reporting status indicators, $\mathbf{X}_{tci} = \begin{pmatrix}X_{tci}^{PR} & X_{tci}^{LR} & X_{tci}^{NR}\end{pmatrix}^T$, can be assumed to follow a point-multinomial distribution

$$\mathbf{X}_{tci} \sim Multinomial\left(1, p_{PRci}, p_{LRci}, p_{NRci}\right)$$

$$X_{tci}^{l} = 0,1 \text{ (l=PR, LR, NR)}$$

$$\sum_{l=PR,LR,NR} X_{tci}^{l} = 1$$

$$p_{lci} \in \left[0,1\right], \sum_{l} p_{lci} = 1, \; \left(l = \text{PR,LR,NR}\right)$$

For a sample unit that is not preliminary reporter, there are two states that can occur – late reporting or nonresponse. The conditional distribution of reporting status indicators for late reporting and nonresponse, given a sample unit is not a preliminary reporter, can be shown to follow a point binomial distribution

$$\left(X_{tci}^{LR} \quad X_{tci}^{NR}\right)^{T} \mid X_{tci}^{PR} = 0 \sim Bin\left(1, p_{LR|X_{tci}^{PR}=0,ci}, p_{NR|X_{tci}^{PR}=0,ci}\right)$$

$$X_{tci}^{l} = 0{,}1, \text{ (l=LR, NR)}$$

$$\sum_{l=LR,NR} X_{tci}^{l} = 1$$

$$p_{l|X_{tci}^{PR}=0,ci} \in [0,1], \sum_{l} p_{l|X_{tci}^{PR}=0,ci} = 1, \ (l = \text{LR,NR})$$

A logit model is thus appropriate to describe the conditional probability a sample unit is a late reporter in month t $\left(X_{tci}^{LR} = 1\right)$ for a sample unit, given the sample unit is not a preliminary reporter in month t $\left(X_{tci}^{PR} = 0\right)$

$$\text{logit}\left[P\left(X_{tci}^{LR} = 1 \mid X_{tci}^{PR} = 0, \mathbf{\Psi}\right)\right] = \alpha_{\text{c}} + \boldsymbol{\gamma}_{c}^{T}\mathbf{\Psi}_{tci}$$

where

$\dfrac{\exp(\alpha_c)}{1+\exp(\alpha_c)}$ (i.e., inverse logit$[\alpha_{\text{c}}]$) represents the underlying cell-level conditional probability of late reporting

$\mathbf{\Psi}_{tci}$ is the vector of factor values for the sample unit

$\boldsymbol{\gamma}_{c}$ is the vector of factor coefficients

Review of CES reporting patterns earlier in this chapter suggests factors be defined by the following characteristics representing respondent, operational, and environmental factors:

*Number of months (through month $t-1$) since last report (G=0, 1, 2, …)

*Reporting status the prior two months $\mathbf{X}_{(t-1)ci}, \mathbf{X}_{(t-2)ci}$

*Design size class (S<10, 10-19, 20-49, 50-99, 100-249, 250-499, 500-999, 1000+)

*Length of pay period (L=Weekly, Bi-weekly, Semi-monthly, monthly)

*Number of reporting days for the month (D=9, 10, 11, 12, 13, 14, 15)

Based on results of the reporting pattern profile, number of months from last report was translated to

$$\ln(G+1)$$

for use in the model.

All factors are categorical. As categories were nominal, each factor was translated to a vector of dummy $(0,1)$ values for use in the model. Each vector contains all 0's and a single 1 to designate the factor category for the sample unit. (This has already been done for prior reporting status.) For example, the length of pay period categories are translated to the following vector

$$\mathbf{L}_{ci} = \left(L_{(\text{Weekly})ci} \quad L_{(\text{Bi-weekly})ci} \quad L_{(\text{Semi-Monthly})ci} \quad L_{(\text{Monthly})ci}\right)$$

$L_{(lopp)ci} = 0,1$; lopp=Weekly, Bi-weekly, Semi-monthly, Monthly

A seasonal component, corresponding to a calendar month effect, may also be present in the underlying model. However, the number of months available for the research (15 months at the outset, growing to only 26 months) was deemed insufficient to allow this effect to either be estimated or to be distinguished from the number of reporting days effect. This constraint, in conjunction with results of the reporting pattern profile led to exclusion of a seasonal component for the working model.

A collection mode effect may also be present in the underlying model. Research by Rosen, et al. (1993) indicates differential rates of reporting by collection mode. There may be a similar effect related to conditional probability of late reporting as well. Intuitively, one could posit such an effect for mail vs. automated forms of collection due to the delay associated with mail delivery. Collection mode effect was initially planned to be included in the dissertation research; however, obstacles to data availability were encountered due to the current status of reporting and retention of information on collection mode, resulting in lack of complete, accurate information covering all months and all CES sample units. As a result, collection mode was excluded in the working model. Extension of the working model to include collection mode, if complete and accurate data could be obtained, could yield additional predictive power.

For purposes of estimation, a Bayes approach was used. This approach was used rather than logistic regression estimation as the number of parameters involved in the model resulted in sparse or missing cells, and Bayes estimation provides parameter estimates for such situations. The working model associated with $X_{tci}^{LR} \mid X_{tci}^{PR} = 0$ (reporting status of late reporting, given a sample unit is not a preliminary reporter) was formulated as point-binomial distribution, with binomial probability following a logit model and factor coefficients of the logit model assumed to have uniform priors

$$X_{tci}^{LR} \mid X_{tci}^{PR} = 0, \mathbf{X}_{(t-1)ci}, \mathbf{X}_{(t-2)ci}, G_{(t-1)ci}, \mathbf{S}_{ci}, \mathbf{L}_{ci}, \mathbf{D}_{t}, \mathbf{\Phi}_{tc}^{LR} \sim Bin\left(1, p_{LR|X_{tci}^{PR}=0,ci}\right)$$

$$\operatorname{logit}\left(p_{LR|X^{PR}_{tci}=0,ci}\right)=$$

$$\alpha_{\mathrm{c}}+\boldsymbol{\gamma}^{T}_{(\mathrm{t}\text{-}1)c}\mathbf{X}_{(t-1)ci}+\boldsymbol{\gamma}^{T}_{(\mathrm{t}\text{-}2)c}\mathbf{X}_{(t-2)ci}+\gamma_{\mathrm{Gc}}\ln\left(G_{(t-1)ci}+1\right)+\boldsymbol{\gamma}^{T}_{\mathrm{Sc}}\mathbf{S}_{ci}+\boldsymbol{\gamma}^{T}_{\mathrm{Lc}}\mathbf{L}_{ci}+\boldsymbol{\gamma}^{T}_{\mathrm{Dc}}\mathbf{D}_{t}$$

$$\mathbf{X}_{(t-k)ci}=\begin{pmatrix}X^{PR}_{(t-k)ci}\\ X^{LR}_{(t-k)ci}\\ X^{NR}_{(t-k)ci}\end{pmatrix}\text{, k=1,2}$$

$G_{(t-1)ci}=(0,1,2,...)$ is the number of months (through month $t-1$) since sample unit i in estimation cell c last reported

$\mathbf{S}_{ci}=\begin{pmatrix}S_{1ci}\\ \vdots\\ S_{8ci}\end{pmatrix}$ is the vector of dummy size variables

$\mathbf{L}_{ci}=\begin{pmatrix}L_{(Weekly)ci}\\ \vdots\\ L_{(Monthly)ci}\end{pmatrix}$ is the vector of dummy length of pay period variables

$\mathbf{D}_{t}=\begin{pmatrix}D_{9t}\\ \vdots\\ D_{15t}\end{pmatrix}$ is the vector of dummy number of reporting days variables

$$\boldsymbol{\Phi}^{LR}_{tc}=\left(\alpha_{c},\boldsymbol{\gamma}^{T}_{(\mathrm{t}\text{-}1)c},\boldsymbol{\gamma}^{T}_{(\mathrm{t}\text{-}2)c},\gamma_{\mathrm{Gc}},\boldsymbol{\gamma}^{T}_{\mathrm{Sc}},\boldsymbol{\gamma}^{T}_{\mathrm{Lc}},\boldsymbol{\gamma}^{T}_{\mathrm{Dc}}\right)$$

$$\boldsymbol{\gamma}_{(t-1)c}=\begin{pmatrix}\gamma^{PR}_{(t-1)c}\\ \gamma^{LR}_{(t-1)c}\\ \gamma^{NR}_{(t-1)c}\end{pmatrix},\ \boldsymbol{\gamma}_{(t-2)c}=\begin{pmatrix}\gamma^{PR}_{(t-2)c}\\ \gamma^{LR}_{(t-2)c}\\ \gamma^{NR}_{(t-2)c}\end{pmatrix}$$

$$\boldsymbol{\gamma}_{Sc}=\begin{pmatrix}\gamma_{S_1c}\\ \vdots\\ \gamma_{S_8c}\end{pmatrix},\ \boldsymbol{\gamma}_{Sc}=\begin{pmatrix}\gamma_{L_{(Weekly)}c}\\ \vdots\\ \gamma_{L_{(Monthly)}c}\end{pmatrix},\ \boldsymbol{\gamma}_{D}=\begin{pmatrix}\gamma_{D_9}\\ \vdots\\ \gamma_{D_{15}}\end{pmatrix}$$

$$\alpha_{c}\sim U\left(l_{\alpha},u_{\alpha}\right)$$

$$\gamma_{(t-k)c}^{l} \sim U\left(-b_X, b_X\right),\ \left(l = \text{PR,LR,NR}\right),\ \left(k = 1, 2\right)$$

$$\gamma_{Gc} \sim U\left(-b_G, b_G\right)$$

$$\gamma_{S_k c} \sim U\left(-b_S, b_S\right),\ \left(k = 1, \ldots, 8\right)$$

$$\gamma_{L_k c} \sim U\left(-b_L, b_L\right),\ \left(k = 1, \ldots, 4\right)$$

$$\gamma_{D_k c} \sim U\left(-b_D, b_D\right),\ \left(k = 1, \ldots, 7\right)$$

$l_\alpha, u_\alpha, b_X, b_G, b_S, b_L, b_D$ are pre-defined bounds for the corresponding prior distributions, defined using the following assumptions

Underlying cell-level conditional probabilities of late reporting range between .01 and .99, thus $\left(l_\alpha, u_\alpha\right) = \left(-2, 2\right)$

Effect due to gap in reporting and prior months' reporting status are expected to be greater than effect due to other categorical factors (design size class, length of pay period, number of reporting days), with the following bounds selected

$$\left(-b_X, b_X\right) = \left(-5, 5\right)$$

$$\left(-b_G, b_G\right) = \left(-5, 5\right)$$

$$\left(-b_S, b_S\right) = \left(-2, 2\right)$$

$$\left(-b_L, b_L\right) = \left(-2, 2\right)$$

$$\left(-b_D, b_D\right) = \left(-2, 2\right)$$

In practice, the dimension for each vector was reduced by one, since any one element is linearly dependent on the remaining elements (as $\sum \gamma_k = 1$). The element selected for exclusion from the vector becomes the reference level for the factor. The reference level for each categorical factor was selected as follows

Reporting status: $X^{PR}_{(t-k)ci} = 1, \ (k = 1, 2)$

Design size class: S=50-99

Length of pay period: L=Bi-weekly

Number of reporting days: D=12

For reporting status, preliminary reporting was designated as the reference level. For each of the remaining variables, the level roughly in the middle of the range of levels was designated as the reference level.

The model was further refined in an attempt to reduce the number of parameters in the model. Collapsing of categories within a factor was carried out on the basis of estimated values for the factor coefficients obtained using the full set of months available for the research (March 2000 through December 2002, as described in Chapter III) using WinBUGS v1.4. WinBUGS was called from a program written in R v.1.8.1, using background code developed by Andrew Gelman (see Gelman, et al., 2003). (Note: For Manufacturing, the sample file exceeded allowable space limits for the software. Approximately half the sample was randomly selected within each month for use in the modeling. Each observation was assigned a random number generated using a standard normal distribution. Observations with random numbers greater than zero were selected.)

The following discussion is based on information contained in Sinharay (2003) MCMC algorithms, such as those used within WinBUGS, are used to obtain a random sample from a posterior distribution of interest given sample data and prior distributions. This sample is used to approximate the posterior distribution, allowing posterior expectations for parameters associated with the distribution to be derived (see, e.g., Gelman, et al., 1995). This process is carried out by specifying prior distributions for the variable of interest and parameters of the distribution, along with sample data. The user also specifies the number of iterations to be run by the algorithm, a burn-in period (the number of initial iterations discarded), and a number of chains to be run (the number of separate series of iterations run). Finally, the user specifies initial values for the parameters of the distribution.

The MCMC algorithm seeks to create a distribution that has converged to the posterior distribution of interest. Gelman and Rubin (1992) proposed a "potential scale reduction factor" (PSRF) as an estimate of how much sharper the distribution estimate might become if the simulations were continued indefinitely. This PSRF declines to one as the simulated distribution converges to the posterior distribution. Generally, values of PSRF less than 1.1 or 1.2 are acceptable.

The WinBUGS software offers several additional options which are useful in checking for convergence. The first is to run multiple chains, that is, to create multiple initial values and sets of simulations to see that they converge to the same estimates. The second is to set the number of iterations run by the MCMC algorithm sufficiently high so as to achieve some level of convergence.

Model diagnostics provided through WinBUGS are the Deviance Information Criterion (DIC, intended as a generalization of Akaike's Information Criteria) and p(D) (effective number of parameters) (Spiegelhalter, et al. 2002). These diagnostics are used in comparing competing models.

Output generated from WinBUGS in R under the Gelman software includes the following: 1) parameter estimates, standard deviations, and selected percentiles; 2) values of PSRF values for each parameter; 3) graphs of parameter estimates and PSRF values. Code can be added to the R program to capture DIC and pD values for the model.

The full model was run using two chains, with 1,000 iterations and a burn-in period of 500 iterations. Initial values for each parameter were set at 0.1 above the mean for the distribution for chain one and 0.1 below the mean for the distribution for chain two. Table 22 contains PSRF values for the various parameters by industry. As can be seen, the only parameter which did not meet the guideline convergence criteria is the intercept for Manufacturing (PSRF=1.27), which exceeds the criteria only slightly. Based on this information, the model was not run using a larger number of iterations. Appendix D contains factor estimates and associated standard deviations from the full model.

Table 22-PSRF Values for Full Conditional Reporting Status Model

Logit Model for Conditional Probability of Late Reporting
Potential Scale Reduction Factors

		Construction	Manufacturing	Mining	Wholesale Trade
Intercept		1.04	1.27	1.00	1.16
Number of Reporting Days	9	1.00	1.05	1.01	1.03
	10	1.00	1.06	1.01	1.05
	11	1.00	1.04	1.01	1.04
	12*	n/a	n/a	n/a	n/a
	13	1.00	1.04	1.01	1.03
	14	1.01	1.05	1.02	1.05
	15	1.01	1.01	1.01	1.02
Length of Pay Period	Weekly	1.03	1.01	1.01	1.00
	Bi-Weekly*	n/a	n/a	n/a	n/a
	Semi-Monthly	1.00	1.01	1.01	1.00
	Monthly	1.01	1.00	1.00	1.00
Design Size Class	<10	1.01	1.04	1.01	1.07
	10-19	1.00	1.03	1.00	1.04
	20-49	1.00	1.06	1.00	1.05
	50-99*	n/a	n/a	n/a	n/a
	100-249	1.00	1.11	1.01	1.09
	250-499	1.00	1.13	1.01	1.04
	500-999	1.01	1.12	1.01	1.04
	1000+	1.01	1.06	1.01	1.07
Reporting Pattern	$X^{PR}_{(t-1)ci} = 1$	n/a	n/a	n/a	n/a
	$X^{LR}_{(t-1)ci} = 1$	1.00	1.05	1.01	1.01
	$X^{NR}_{(t-1)ci} = 1$	1.00	1.01	1.00	1.00
	$X^{PR}_{(t-2)ci} = 1$	n/a	n/a	n/a	n/a
	$X^{LR}_{(t-2)ci} = 1$	1.01	1.01	1.16	1.02
	$X^{NR}_{(t-2)ci} = 1$	1.00	1.03	1.02	1.00
ln(Reporting Gap)		1.00	1.02	1.12	1.03

*Designated reference level for factor

For illustration purposes, model results generated by the R program are displayed graphically in Figure 13 for Mining and Figure 14 for Wholesale Trade. By way of explanation of the figures, the graph on the left shows the posterior 80% interval for each parameter along with the PSRF value (designated as R-hat in the graph). The parameter "a" corresponds to the intercept, "gD[k]" corresponds to coefficient for the kth level of the number of reporting days indicator, "gL[k]" corresponds to coefficient for the kth level of the length of pay period indicator, "gS[k]" corresponds to coefficient for the kth level of the design size class indicator, “gLRk” corresponds to coefficient for the late reporter indicator for month $t-k$, “gNRk” corresponds to coefficient for the nonreporter indicator for month $t-k$, and “gG” corresponds to the coefficient for the log(1 + reporting gap length) parameter.

The graphs on the right show the posterior median and 80% intervals associated with each of the two chains for each parameter. For parameters with different levels (e.g., design size class, “gS”), each level is provided on the same graph. In addition, there is a graph for the deviance under the model. The nearly monotonic increasing impact of design size class is visible for both industries (i.e., the medians increase or are stable from level to level within “gS”), along with the influence of monthly length of pay period and 15 reporting days.

Figure 13-Conditional LR Rate Model Results, Full Model: Mining

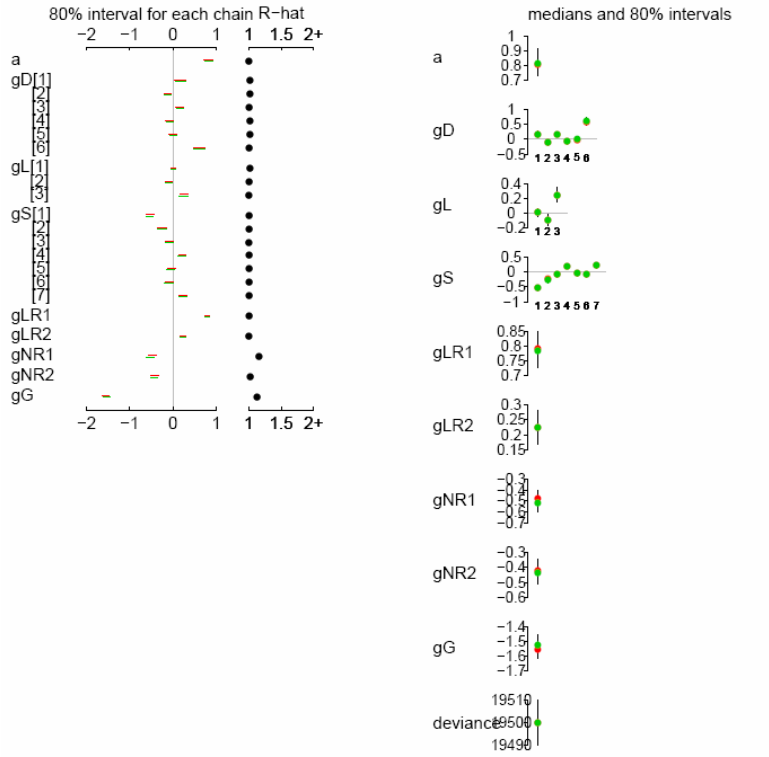
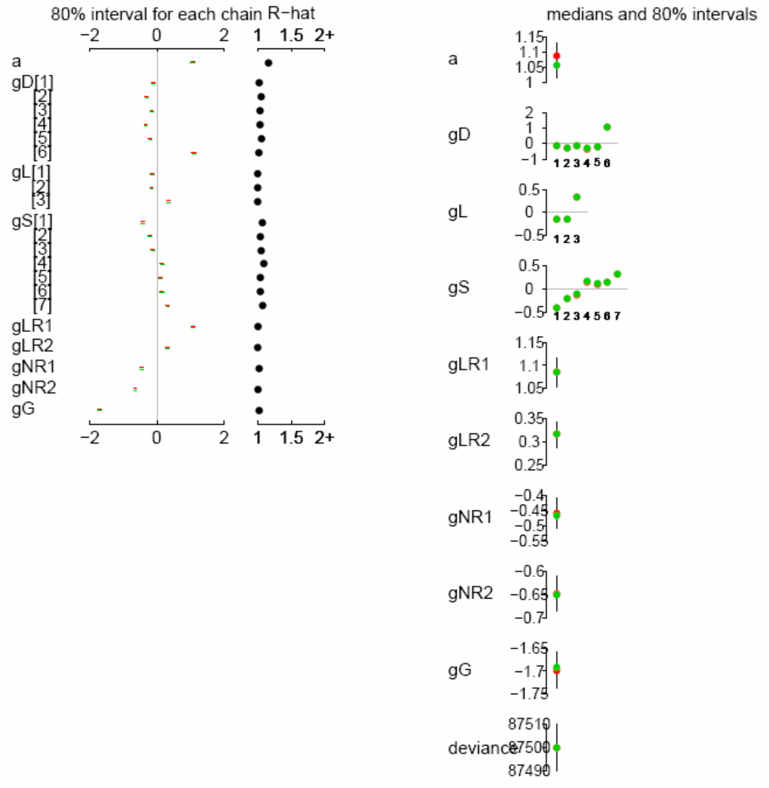


Figure 14-Conditional LR Rate Model Results, Full Model: Wholesale Trade



In the interest of parsimony of the model and reduction of computer calculation time to run the model, levels within a categorical variable were collapsed if posterior 95% credible intervals for the coefficient for one category encompassed the estimated coefficient for another category. Collapsed factor categories selected for each industry are provided in Table 23.

Collapsing follows the expected relationship among design size classes, wherein the likelihood of late reporting increases as design size class increases, with the exception of Mining. This exception may be due in part to smaller sample sizes for Mining, especially for larger size classes. Collapsing among number of reporting data does not follow an intuitive pattern (e.g., late reporting in Construction associated with 10 reporting days less similar to 9 or 11 reporting days than to 13 reporting days). However, as discussed previously, this may be an indication of a more complex relationship involving reporting timeframe and length of pay period.

Table 23-Collapsed Factor Categories for Logit Model of Conditional LR Rate

Logit Model for Conditional Probability of Late Reporting
Collapsed Factor Categories

		Construction	Manufacturing	Mining	Wholesale Trade
Number of Reporting Days	9	D1	D1	D1	D1
	10	D2	D0	D0	D2
	11	D3	D1	D1	D1
	12*	D0	D0	D0	D0
	13	D2	D0	D0	D2
	14	D0	D0	D0	D1
	15	D0	D0	D2	D3
Length of Pay Period	Weekly	L1	L1	L0	L1
	Bi-Weekly*	L0	L0	L0	L0
	Semi-Monthly	L2	L1	L0	L1
	Monthly	L3	L2	L1	L2
Design Size Class	<10	S1	S1	S1	S1
	10-19	S2	S1	S2	S2
	20-49	S3	S2	S0	S3
	50-99*	S0	S0	S0	S0
	100-249	S0	S3	S3	S4
	250-499	S0	S3	S0	S4
	500-999	S0	S3	S0	S4
	1000+	S0	S4	S3	S5
Reporting Pattern	$X^{PR}_{(t-1)ci}=1$	R(t-1)0	R(t-1)0	R(t-1)0	R(t-1)0
	$X^{LR}_{(t-1)ci}=1$	R(t-1)1	R(t-1)1	R(t-1)1	R(t-1)1
	$X^{NR}_{(t-1)ci}=1$	R(t-1)2	R(t-1)2	R(t-1)2	R(t-1)2
	$X^{PR}_{(t-2)ci}=1$	R(t-2)0	R(t-2)0	R(t-2)0	R(t-2)0
	$X^{LR}_{(t-2)ci}=1$	R(t-2)1	R(t-2)1	R(t-2)1	R(t-2)1
	$X^{NR}_{(t-2)ci}=1$	R(t-2)2	R(t-2)2	R(t-2)2	R(t-2)2

****F*** 0 represents reference level for factor ***F***

To assess the appropriateness of proposed collapsing, the model was run with and without collapsing for the beginning month of the period of interest for the research, April, 2001. Model diagnostics DIC and p(D) under the two approaches are provided in Table 24. The DIC values indicate the model with collapsed factors experiences no noticeable loss of information from the full model. Therefore, these reduced sets of models were used in the empirical analysis.

Table 24-Reporting Status Model: Diagnostics for Full, Collapsed Set of Parameters

Model Diagnostics
Reporting Status Model
Based on March 2000 - March 2001 Reporting History

	Full set of parameters		Collapsed set of parameters	
	DIC	pD	DIC	pD
Construction	51650	29.0	51652	18.2
Manufacturing	43040	26.6	43136	17.1
Mining	7408	23.0	7422	11.2
Wholesale Trade	34915	25.8	34979	21.7

For the empirical analysis, no variance estimates were calculated. Notes concerning variance estimation for the reporting status model are provided in Appendix E.

E. Model Implementation

1. Approach

Reporting status likelihoods for sample units not reporting as of the preliminary cutoff date for month t were estimated using conditional probabilities of late reporting resulting from the model for the period April 2001 through March 2002. Estimated conditional late reporting rates for each month were compared to actual values.

2. Generating Estimates

Parameter estimates were generated for each month of interest, t, using the model in conjunction with all available data from January 2000 through month $t-1$. Parameters for the logit model for the conditional probability of late reporting status were estimated using WinBUGS v1.4 called from a program written in R v.1.8.1.

The WinBUGS model specification is provided in Appendix F.1. The R code used for parameter estimation is provided in Appendix F.2.

The model was run using two chains, with 500 iterations and a burn-in period of 250 iterations. Initial values for each parameter were set at 0.1 above the mean for the distribution for chain one and 0.1 below the mean for the distribution for chain two. Averages for the potential scale reduction factors for the model across the 12 months are provided in Table 25. As can be seen, there were 15 parameters that failed to meet the guideline convergence criteria for at least one month. Further examination showed failure occurred in just one month for all but three parameters (Construction, ln(Reporting Gap+1) – 2 months, Wholesale Trade, Design size class 3 – 3 months, and Wholesale Trade, intercept – 5 months). As maximum PSRF values the parameters with multiple occurrences were not dramatically greater than 1.2 and the other parameters had at most one occurrence, the model was not run using additional iterations. The lack of convergence for a given month could, however, adversely affect the predictive power of the model.

Table 25-PSRF Values for Conditional Reporting Status Model

Logit Model for Conditional Probability of Late Reporting
Maximum Potential Scale Reduction Factors
April 2001 - March 2002

		Construction	Manufacturing	Mining	Wholesale Trade
Intercept		1.42	1.40	1.07	1.32
Number of Reporting Days	D[1]	1.03	1.02	1.03	1.16
	D[2]	1.05	0.00	1.09	1.12
	D[3]	1.03	n/a	n/a	1.06
Length of Pay Period	L[1]	1.12	1.02	1.02	1.02
	L[2]	1.53	1.14	n/a	1.03
	L[3]	1.16	n/a	n/a	n/a
Design Size Class	S[1]	1.03	1.15	1.01	1.35
	S[2]	1.02	1.15	1.01	1.18
	S[3]	1.02	1.38	1.06	1.26
	S[4]	n/a	1.24	n/a	1.38
	S[5]	n/a	n/a	n/a	1.28
	$X^{LR}_{(t-1)ci}=1$	1.03	1.02	1.09	1.08
	$X^{NR}_{(t-1)ci}=1$	1.07	1.05	1.03	1.03
	$X^{LR}_{(t-2)ci}=1$	1.22	1.65	1.19	1.24
	$X^{NR}_{(t-2)ci}=1$	1.07	1.21	1.09	1.02
ln(Reporting Gap)		1.27	1.64	1.18	1.16

Several illustrations from the graphical results available from the R software used to call WinBUGS are provided in Figure 15-Figure 18.

Figure 15-Conditional LR Rate Model Results for March 2002: Construction

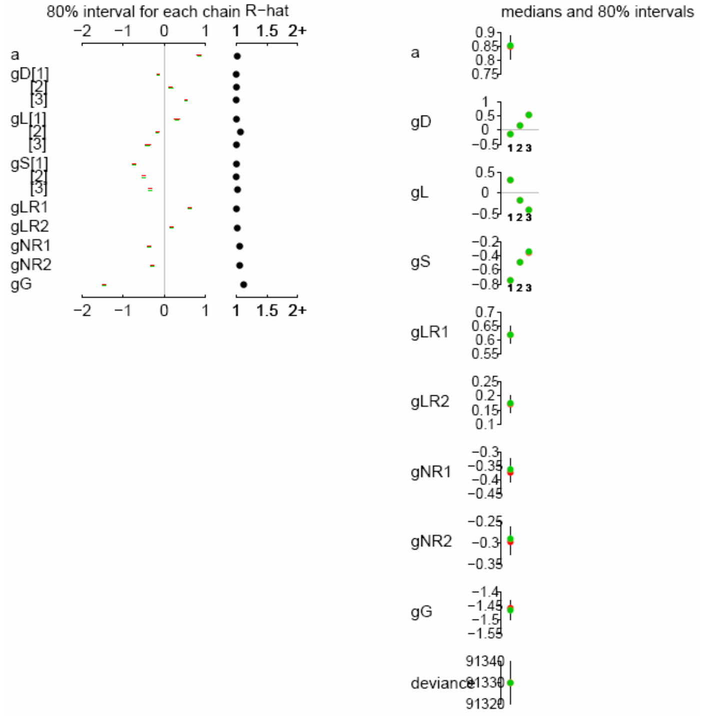


Figure 16- Conditional LR Rate Model Results for March 2002: Manufacturing

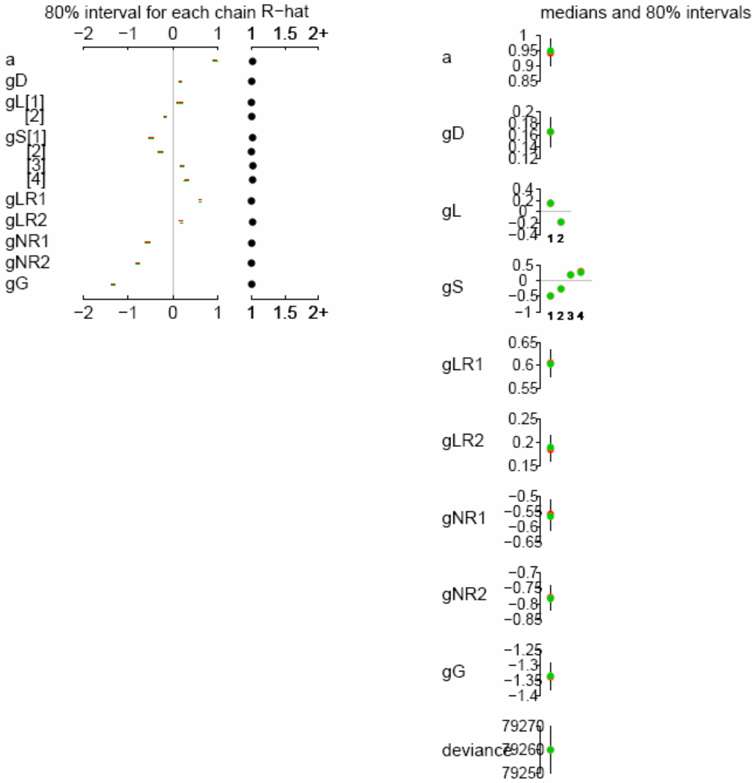


Figure 17- Conditional LR Rate Model Results for March 2002: Mining

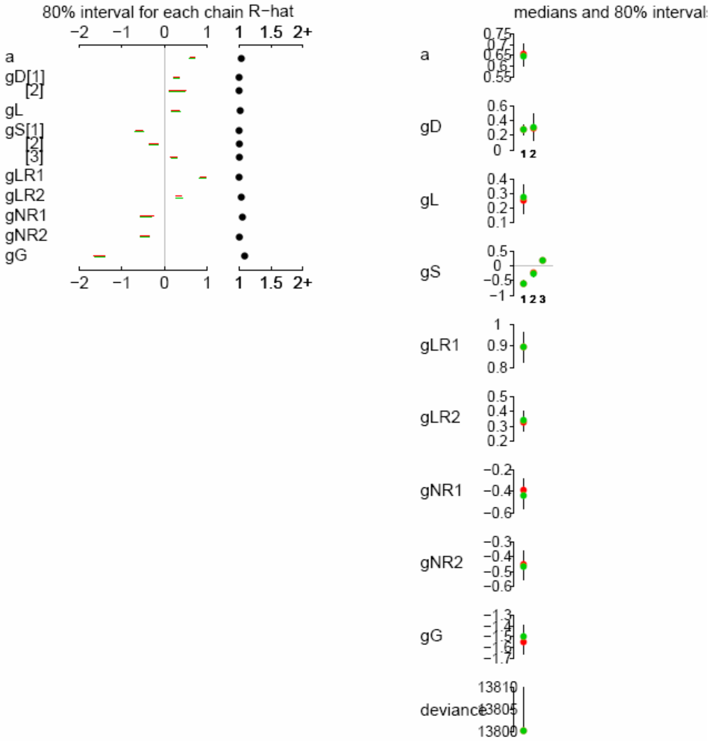
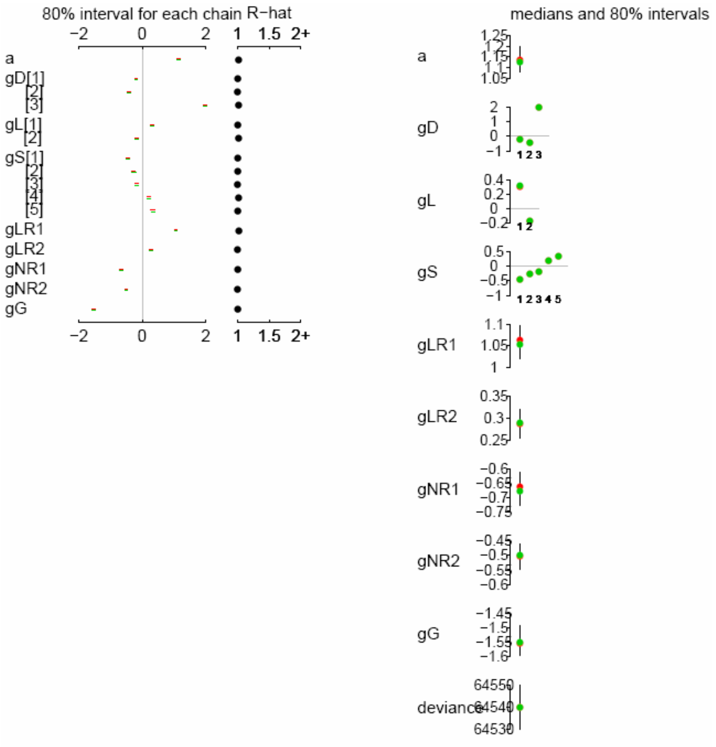


Figure 18- Conditional LR Rate Model Results for March 2002: Wholesale Trade



Estimated coefficient values for the initial month of the analysis period, April 2001, are provided in Table 26. It should be remembered that factor level definitions vary across industry for number of reporting days, design size class, and, with the exception of L1 (Monthly), length of pay period.

Table 26-Coefficient Estimates for Conditional Late Reporting Model: April 2001

Logit Model for Conditional Probability of Late Reporting
Coefficient Estimates-Collapsed Model
April 2001

		Construction			Manufacturing			Mining			Wholesale Trade		
		2.5% Level	Estimate	97.5% Level	2.5% Level	Estimate	97.5% Level	2.5% Level	Estimate	97.5% Level	2.5% Level	Estimate	97.5% Level
Intercept		0.6212	0.7151	0.8001	0.8815	0.9590	1.0391	0.3475	0.4719	0.5885	1.2175	1.3181	1.4185
Number of Reportin g Days	D1	-0.2264	-0.1801	-0.1311	0.1454	0.1973	0.2528	0.3686	0.5035	0.6567	-0.4646	-0.3905	-0.3221
	D2	0.2929	0.3787	0.4601	n/a	n/a	n/a	0.1186	0.4078	0.6584	-0.6097	-0.5338	-0.4617
	D3	0.9263	0.9853	1.0480	n/a	n/a	n/a	n/a	n/a	n/a	1.8379	1.9383	1.9975
Length of Pay Period	L1	0.1401	0.2469	0.3462	-0.0287	0.0923	0.2103	0.1053	0.3194	0.5357	0.1451	0.2420	0.3399
	L2	-0.3462	-0.2809	-0.2111	-0.2576	-0.2077	-0.1479	n/a	n/a	n/a	-0.3339	-0.2780	-0.2179
	L3	-0.5977	-0.5057	-0.4060	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Design Size Class	S1	-0.7410	-0.6842	-0.6288	-0.6006	-0.4908	-0.4016	-0.8320	-0.6297	-0.4287	-0.4764	-0.3704	-0.2707
	S2	-0.5391	-0.4750	-0.4079	-0.4387	-0.3238	-0.2235	-0.6391	-0.4325	-0.2295	-0.3485	-0.2212	-0.1009
	S3	-0.4090	-0.3548	-0.2915	0.1516	0.2157	0.2890	0.0608	0.1968	0.3274	-0.2181	-0.1067	-0.0071
	S4	n/a	n/a	n/a	0.2730	0.3538	0.4317	n/a	n/a	n/a	0.1061	0.1942	0.2865
	S5	n/a	n/a	n/a	n/a	n/a	n/a	mn/a	mn/a	mn/a	0.3236	0.4143	0.5053
Reporting Pattern	$X_{(t-1)ci}^{LR} = 1$	0.6087	0.6674	0.7283	0.5338	0.5922	0.6524	0.7844	0.9280	1.0760	0.9873	1.0539	1.1226
	$X_{(t-1)ci}^{NR} = 1$	0.1501	0.2055	0.2573	0.1224	0.1780	0.2354	0.3962	0.5184	0.6501	0.1725	0.2314	0.2979
	$X_{(t-2)ci}^{LR} = 1$	-0.4928	-0.4023	-0.2855	-0.9884	-0.8760	-0.7664	-0.9716	-0.6470	-0.3858	-1.0325	-0.8898	-0.7592
	$X_{(t-2)ci}^{NR} = 1$	-0.2124	-0.1485	-0.0812	-0.8647	-0.7811	-0.7105	-0.3813	-0.1946	-0.0086	-0.4985	-0.4114	-0.3235
ln(Reporting Gap)		-1.4035	-1.3175	-1.2409	-1.0605	-0.9533	-0.8587	-1.4816	-1.2337	-0.9883	-1.3750	-1.2515	-1.1105

The reporting gap has a large negative effect on the conditional late reporting rate, yielding an expected decline of 31 to 39 percentage points in the late reporting rate across industries, due to a change from no reporting gap to a gap of one month for a sample unit with characteristics corresponding to the reference levels for the remaining factors. In the other direction, prior month late reporting status for a sample unit with characteristics corresponding to the reference levels for the remaining factors is associated with an expected increase of 10 to 19 percentage points in the likelihood of current month late reporting.

There were some shifts in the values of the estimated parameters across time, as indicated in Figure 19-Figure 23. The estimated coefficient for log of reporting gap decreased between April 2001 and March 2002. This was somewhat offset for Manufacturing and Wholesale Trade by an increase in the estimated coefficient for prior month nonresponse. As a reporting gap of one or more month implies prior month nonresponse, this suggests an interaction term for the model could be considered in future research. The influence of prior month late reporting, by contrast, was relatively stable across the analysis time period. The intercept was also fairly stable across time.

Figure 19-Coefficient Estimates for Log(Reporting Gap+1)

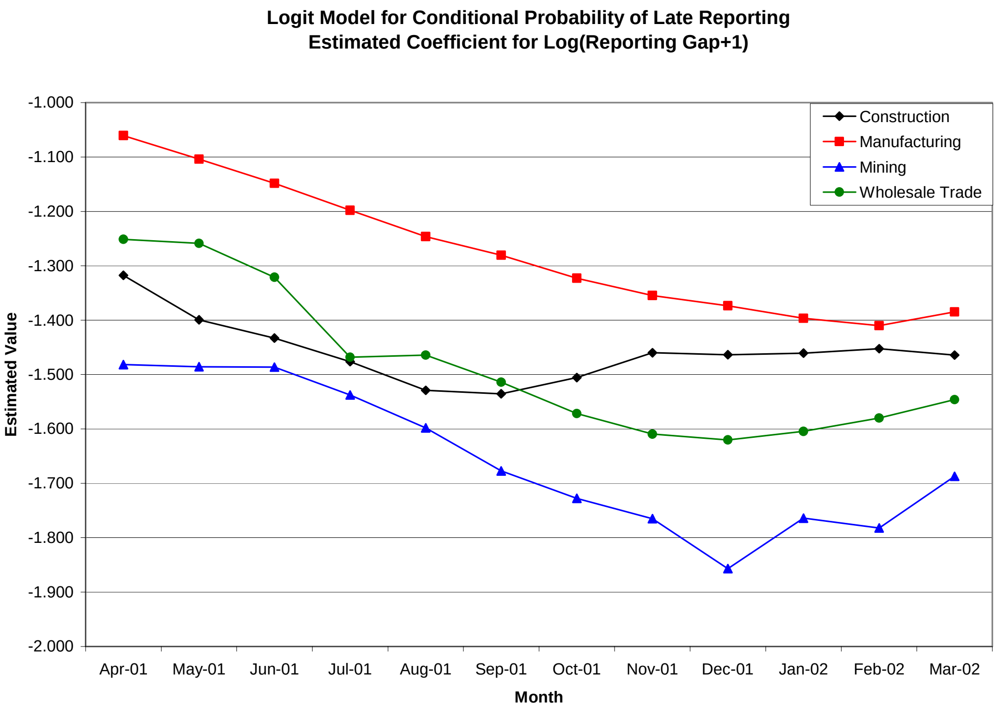


Figure 20-Coefficient Estimates for Prior Month NR

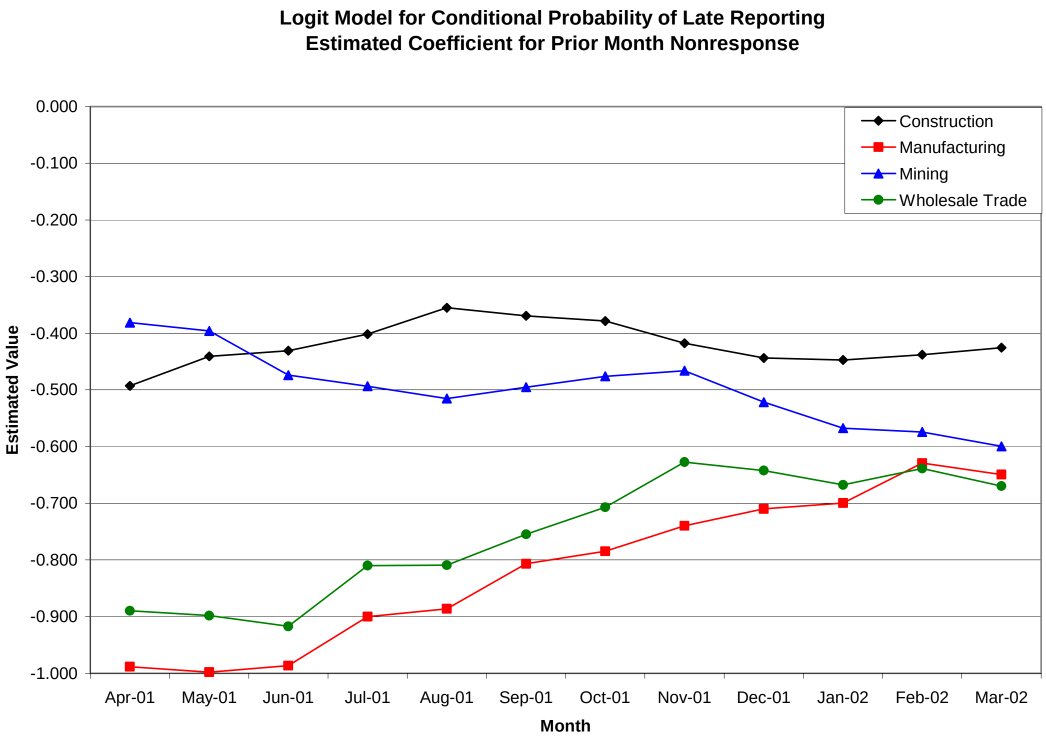


Figure 21-Coefficient Estimates for Prior Month LR

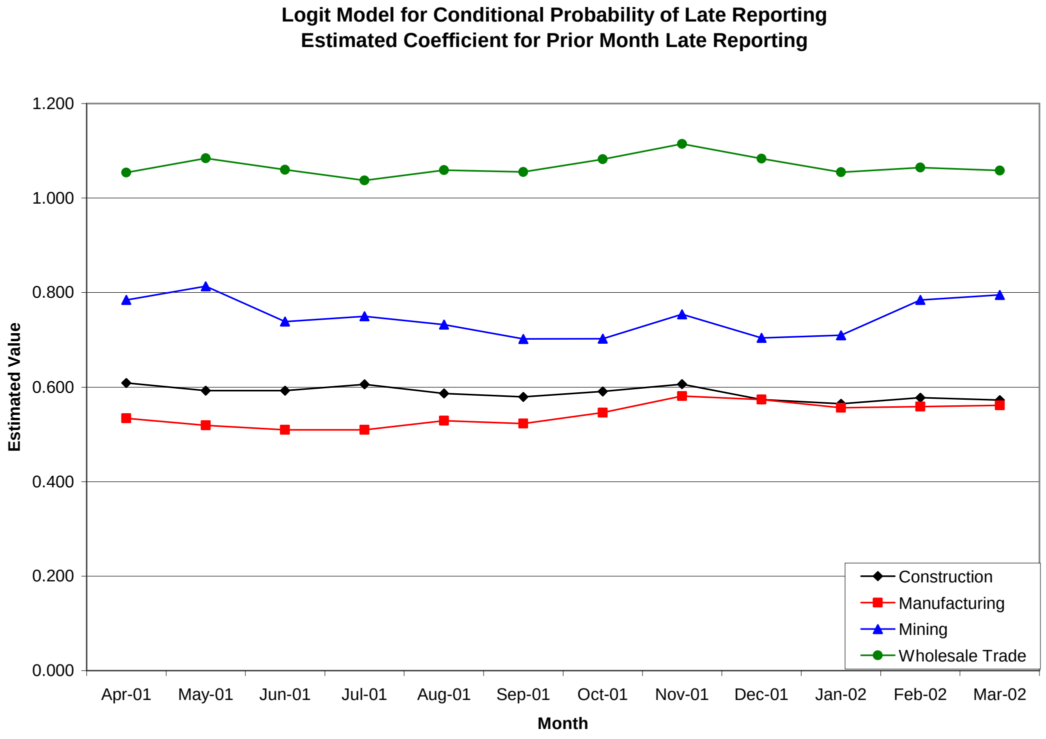


Figure 22-Coefficient Estimates for Length of Pay Period=Monthly

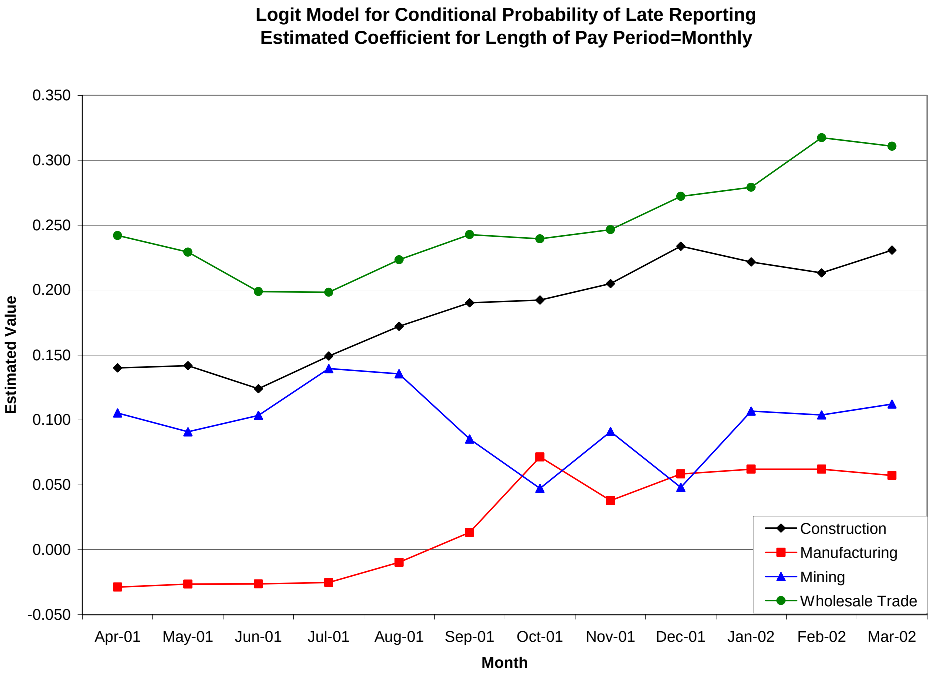
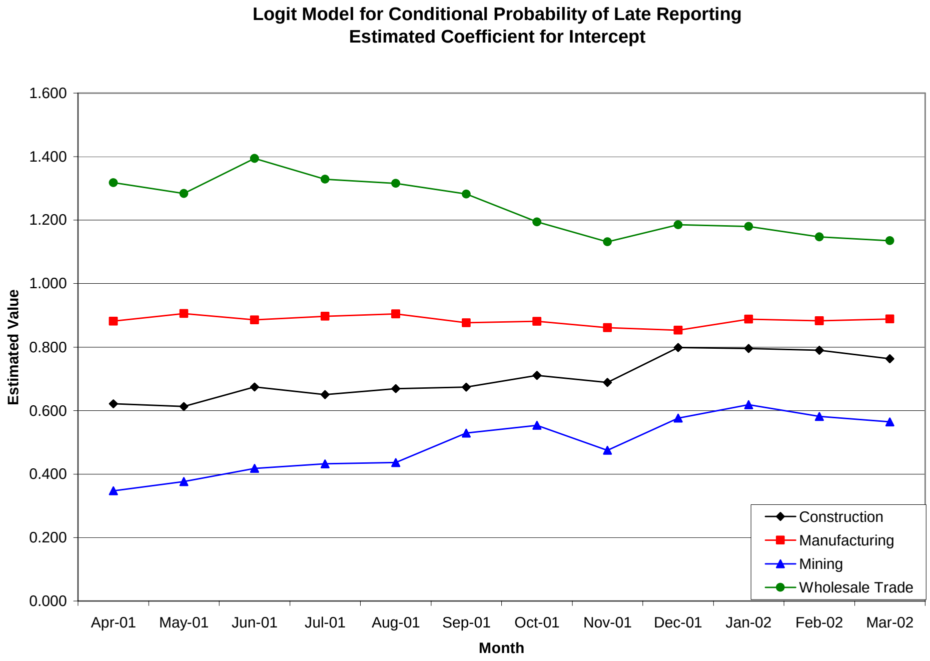


Figure 23-Intercept Estimates



Actual conditional late reporting rates were derived for the period from April 2001 through March 2002, using the revised CES datafile. Estimated conditional late reporting rates were derived using Model III in conjunction with the parameter estimates. Two estimated conditional late reporting rates were derived – using parameter estimates based upon all available data as of month t (updated parameters), and using parameter estimates based upon all available data as of the first month of interest (April '01 parameters). The SAS code used for deriving estimated conditional late reporting rates is provided in Appendix F.3.

3. Measures of Accuracy

The reporting status model was developed to allow accurate prediction of final reporting status for sample units that were not preliminary responders. Assessment of the performance of the model can be made by comparison to actual final reporting status. The measures of accuracy utilized are

$$\mathrm{Err}\left(Est\left(LR_t\right)\right) = Est\left(LR_t\right) - Act\left(LR_t\right)$$

$$\text{Ave Err}\left(\hat{Y}_t^{(0)}\right) = \frac{\sum_{t=Apr'01}^{Mar'02} \mathrm{Err}\left(Est\left(LR_t\right)\right)}{12}$$

$$\text{Ave Abs Err}\left(\hat{Y}_t^{(0)}\right) = \frac{\sum_{t=Apr'01}^{Mar'02} \left|\mathrm{Err}\left(Est\left(LR_t\right)\right)\right|}{12}$$

4. Results

Predicted conditional late reporting rates for the four industries had an average absolute error between two and four percentage points (Table 27-Table 30). There was only one error greater than 10 percentage points, that being the initial month for Construction. Predicted conditional late reporting rates based upon a fixed set of parameter estimates performed almost as well as those based upon updated parameter estimates.

Table 27-Predicted Conditional LR Rates: Construction

Predicted Conditional Late Reporting Rates

April 2001 - March 2002

Construction

Month	Actual LR Rate	Predicted LR Rate (Updated parameters)	Error	Predicted LR Rate (April '01 parameters)	Error
Apr-01	24.8%	40.3%	15.5%	40.3%	15.5%
May-01	29.3%	33.8%	4.5%	34.2%	4.9%
Jun-01	27.0%	26.8%	-0.1%	27.1%	0.2%
Jul-01	27.9%	34.7%	6.8%	38.9%	11.0%
Aug-01	24.6%	24.2%	-0.4%	24.9%	0.3%
Sep-01	24.5%	23.4%	-1.1%	24.1%	-0.4%
Oct-01	24.3%	24.5%	0.3%	25.4%	1.1%
Nov-01	34.6%	32.1%	-2.5%	32.0%	-2.6%
Dec-01	33.2%	25.7%	-7.5%	26.4%	-6.8%
Jan-02	27.1%	27.9%	0.8%	27.1%	0.0%
Feb-02	21.0%	22.1%	1.0%	21.6%	0.6%
Mar-02	20.3%	20.1%	-0.3%	19.8%	-0.3%
Ave Err			1.4%		2.0%
Ave Abs Err			3.4%		3.6%

Table 28-Predicted Conditional LR Rates: Manufacturing

Predicted Conditional Late Reporting Rates

April 2001 - March 2002

Manufacturing

Month	Actual LR Rate	Predicted LR Rate (Updated parameters)	Error	Predicted LR Rate (April '01 parameters)	Error
Apr-01	46.1%	49.2%	3.1%	49.2%	3.1%
May-01	50.6%	50.6%	0.0%	51.4%	0.8%
Jun-01	44.8%	44.1%	-0.7%	44.3%	-0.5%
Jul-01	44.3%	45.8%	1.5%	46.8%	2.5%
Aug-01	39.8%	40.9%	1.1%	41.2%	1.4%
Sep-01	39.4%	40.1%	0.7%	40.6%	1.2%
Oct-01	37.4%	41.8%	4.5%	42.4%	5.0%
Nov-01	40.8%	41.5%	0.7%	42.5%	1.7%
Dec-01	46.3%	41.9%	-4.5%	42.9%	-3.5%
Jan-02	40.8%	42.1%	1.3%	42.7%	1.9%
Feb-02	33.4%	35.1%	1.7%	35.9%	2.6%
Mar-02	32.4%	34.1%	1.8%	35.1%	1.8%
Ave Err			0.9%		1.5%
Ave Abs Err			1.8%		2.2%

Table 29-Predicted Conditional LR Rates: Mining

Predicted Conditional Late Reporting Rates
April 2001 - March 2002
Mining

Month	Actual LR Rate	Predicted LR Rate (Updated parameters)	Error	Predicted LR Rate (April '01 parameters)	Error
Apr-01	41.2%	46.6%	5.3%	46.6%	5.3%
May-01	42.2%	47.6%	5.4%	48.8%	6.6%
Jun-01	39.1%	39.0%	-0.2%	39.2%	0.0%
Jul-01	39.2%	40.4%	1.2%	42.8%	3.6%
Aug-01	45.5%	39.2%	-6.3%	39.3%	-6.2%
Sep-01	41.1%	39.7%	-1.4%	39.7%	-1.4%
Oct-01	28.9%	35.5%	6.6%	34.9%	6.0%
Nov-01	45.9%	37.3%	-8.6%	37.0%	-8.8%
Dec-01	42.8%	38.9%	-3.9%	38.1%	-4.7%
Jan-02	38.0%	43.6%	5.6%	42.6%	4.6%
Feb-02	32.2%	35.1%	2.9%	34.8%	2.6%
Mar-02	24.9%	26.5%	1.6%	27.0%	1.6%
Ave Err			0.7%		0.8%
Ave Abs Err			4.1%		4.3%

Table 30-Predicted Conditional LR Rates: Wholesale Trade

Predicted Conditional Late Reporting Rates
April 2001 - March 2002
Wholesale Trade

Month	Actual LR Rate	Predicted LR Rate (Updated parameters)	Error	Predicted LR Rate (April '01 parameters)	Error
Apr-01	47.1%	48.2%	1.1%	48.2%	1.1%
May-01	53.8%	51.4%	-2.4%	51.6%	-2.3%
Jun-01	47.9%	51.6%	3.7%	51.7%	3.8%
Jul-01	42.7%	45.1%	2.5%	45.3%	2.6%
Aug-01	39.1%	41.5%	2.4%	41.9%	2.8%
Sep-01	37.9%	41.9%	4.0%	43.4%	5.6%
Oct-01	35.9%	39.2%	3.3%	39.5%	3.6%
Nov-01	45.9%	41.9%	-4.0%	42.4%	-3.5%
Dec-01	42.3%	38.9%	-3.5%	39.8%	-2.5%
Jan-02	38.9%	41.6%	2.6%	41.8%	2.9%
Feb-02	33.1%	33.9%	0.8%	34.6%	1.5%
Mar-02	33.0%	32.5%	-0.5%	33.4%	-0.5%
Ave Err			0.8%		1.3%
Ave Abs Err			2.6%		2.7%

Looking at the performance of estimated conditional late reporting rates by prior reporting patterns in Table 31, average absolute errors are below 10 percentage points when sample sizes are above 150.

Table 31-Average Absolute Errors for Predicted Conditional LR Rates

Average Absolute Error in Predicted Conditional Late Reporting Rate
by Prior Reporting Pattern
April 2001 - March 2002

Prior 2 Months' Reporting Pattern		Construction				Manufacturing				Mining				Wholesale Trade			
				Average Absolute Error				Average Absolute Error				Average Absolute Error				Average Absolute Error	
Month t-1	Month t-2	Ave n	Actual LR Rate	Predicted LR Rate (Updated parameters)	Predicted LR Rate (April '01 parameters)	Ave n	Actual LR Rate	Predicted LR Rate (Updated parameters)	Predicted LR Rate (April '01 parameters)	Ave n	Actual LR Rate	Predicted LR Rate (Updated parameters)	Predicted LR Rate (April '01 parameters)	Ave n	Actual LR Rate	Predicted LR Rate (Updated parameters)	Predicted LR Rate (April '01 parameters)
PR	PR	798	65.2%	7.7%	8.8%	1605	72.9%	5.3%	5.3%	115	67.8%	11.2%	10.5%	606	69.5%	8.7%	8.6%
	LR	279	66.4%	4.9%	5.6%	699	73.5%	4.1%	4.0%	51	72.4%	12.1%	12.1%	197	71.5%	6.8%	6.6%
	NR	90	43.0%	15.2%	15.6%	131	41.9%	12.4%	12.3%	10	41.3%	29.9%	29.8%	52	42.4%	21.9%	22.1%
LR	PR	232	78.4%	4.9%	5.4%	621	83.0%	3.4%	3.4%	60	84.3%	7.8%	8.6%	175	81.1%	6.7%	6.3%
	LR	325	81.7%	4.9%	5.0%	1240	87.0%	1.2%	1.2%	114	88.1%	3.0%	2.9%	989	93.5%	2.5%	2.6%
	NR	110	62.6%	12.4%	12.9%	269	64.5%	6.7%	6.8%	20	65.9%	15.0%	15.4%	96	67.7%	13.4%	14.2%
NR	PR	287	24.1%	8.7%	9.5%	487	29.4%	8.6%	8.6%	42	27.1%	9.6%	9.6%	188	22.9%	7.5%	7.5%
	LR	126	36.2%	8.4%	8.8%	310	41.4%	5.4%	5.4%	25	36.0%	17.2%	17.6%	114	35.1%	8.3%	8.6%
	NR	3345	3.3%	2.7%	3.4%	4331	4.8%	1.6%	2.4%	408	3.8%	2.0%	2.6%	2114	3.0%	1.6%	2.3%

5. Discussion

The logit model for conditional late reporting status appears to perform well overall and for larger subsets of the population. The coefficients of the parameters are fairly stable over time, suggesting a periodic update of the estimated values should be sufficient for ongoing prediction. Further research into the relationship between seasonality and number of reporting days could provide improvements to the model. Consideration could also be given to establishing a standard set of factor levels across industry for consistency sake, with the loss in information as per the DIC evaluated for model selection.

The parameter estimates from the model could be used to prioritize resources when targeting nonresponse. Those characteristics associated with lower conditional probability of late reporting should be given higher priority in nonresponse followup. The model could also be used dynamically to estimate level of late reporting expected after preliminary data collection, thereby identifying when the makeup of the non-preliminary reporters are such that low levels of late reporting are expected, allowing special nonresponse followup efforts to be put into place prior to completion of data collection.

Chapter V: An Alternative Approach for the CES Preliminary Estimates of Employment

Within this chapter, an alternate approach for use in CES preliminary estimation of employment is developed and its performance assessed relative to the current methodology. The approach seeks to address potential model misspecification error and involves imputing for missing data in an attempt to predict late reporting values that will be used in revised estimates. The objective is to reduce the difference between preliminary and revised estimates.

Prior to specification of the approach, comments on the current estimator are provided in section A, and the nature of model misspecification error is explored and the CES sample evaluated relative to the potential for model misspecification error in section B. The approach is described in section C and its performance evaluated using historical data in section D.

A. Comments on CES Estimation Methodology

The model that yields the weighted link relative as a maximum likelihood estimator (MLE) is a weighted proportional regression model, in which the current month's value is assumed proportional to the prior month's value (West, et al., 1989), with the proportionality factor assumed to vary by estimation cell, c $\left(=1,\ldots,C\right)$, and month.

Model 0: $Y_{tci} = \rho_{tc} Y_{(t-1)ci} + k_{tci}$

$$k_{tci} \overset{\text{ind}}{\sim} N\left(0, \frac{\sigma_k^2 Y_{(t-1)ci}}{w_{ci}}\right)$$

where ρ_{tc} is the model parameter describing the month t expected growth rate for cell c.

Under this model the maximum likelihood estimator (MLE) for ρ_{tc} is

$$\hat{\rho}_{tc} = \frac{\sum_{i \in s_c} w_{ci} Y_{tci}}{\sum_{i \in s_c} w_{ci} Y_{(t-1)ci}}$$

where s_c represents the sample from estimation cell c. This is the complete response form of the current CES weighted link relative. An estimate of current month employment can be written as

$$\hat{Y}_t = \sum_c \hat{\rho}_{tc} Y_{(t-1)c}$$

In practice, population totals, $Y_{(t-1)c}$, are unknown at the time of estimation, and estimation is complicated by the presence of late reporting and nonresponse. The weighted link relative estimator used for CES is a variant of the MLE taking these situations into account by ignoring late reporting and nonresponse by utilizing only sample units which report data in both months t and $t-1$. Estimated employment is obtained by linking back to the most recently available benchmark totals, $Y_{t_B c}$ (which is assumed to be a fixed quantity), through the monthly weighted link relatives. Thus, using the notation developed in Chapter III, the preliminary estimator for month t may be written as the product of weighted link relatives back to t_B and the benchmark totals.

$$\hat{Y}_t^{(0)} = \sum_{c=1}^{C} \left[\left\{ \frac{\sum_{i \in s_{t,(t-1)|0}} w_{ci} Y_{tci}}{\sum_{i \in s_{t,(t-1)|0}} w_{ci} Y_{(t-1)ci}} \frac{\sum_{i \in s_{(t-1),(t-2)|1}} w_{ci} Y_{(t-1)ci}}{\sum_{i \in s_{(t-1),(t-2)|1}} w_{ci} Y_{(t-2)ci}} \prod_{t^*=t_B+1}^{t-2} \frac{\sum_{i \in s_{t^*,(t^*-1)|2}} w_{ci} Y_{t^*ci}}{\sum_{i \in s_{t^*,(t^*-1)|2}} w_{ci} Y_{(t^*-1)ci}} \right\} Y_{t_B c} \right]$$

$$= \sum_{c=1}^{C}\left[\left\{LR^{(0)}_{t,(t-1)c}\, LR^{(1)}_{(t-1),(t-2)c} \prod_{t^*=t_B+1}^{t-2} LR^{(2)}_{t^*,(t^*-1)c}\right\} Y_{t_B c}\right]$$

Under Model 0 it can be seen that, for all revisions, the expected value of the weighted link relative for month t, conditioned on $\mathbf{Y}_{(t-1)}$, is the month t proportionality factor for the estimation cell.

$$E\left(LR^{(k)}_{t,(t-1)c} \mid \text{Model } 0\right) = E\left[\left(\frac{\sum\limits_{i \in s_{t,(t-1)c|k}} w_{ci} Y_{tci}}{\sum\limits_{i \in s_{t,(t-1)c|k}} w_{ci} Y_{(t-1)ci}}\right) \mid \text{Model } 0\right]$$

$$= \frac{\sum\limits_{i \in s_{t,(t-1)c|k}} w_{ci} E\left(Y_{tci} \mid \text{Model } 0\right)}{\sum\limits_{i \in s_{t,(t-1)c|k}} w_{ci} Y_{(t-1)ci}} = \frac{\sum\limits_{i \in s_{t,(t-1)c|k}} w_{ci} \rho_{tc} Y_{(t-1)ci}}{\sum\limits_{i \in s_{t,(t-1)c|k}} w_{ci} Y_{(t-1)ci}} = \rho_{tc}$$

Correspondingly, the expected value of the estimated employment for month t under Model 0, conditioned on the benchmark population values $\mathbf{Y}_{t_B c}$, is equal to the expected population total for month t. This result is derived through a series of conditional expectations, with conditional expectations taken based on each population total prior to month t.

$$E\left(\hat{Y}^{(k)}_t \mid \text{Model } 0\right) = \sum_c E\left(\left[Y_{t_B c} LR^{(k)}_{t,(t-1)c} LR^{(k+1)}_{(t-1),(t-2)c} \prod_{t^*=t_B+1}^{t-2} LR^{(3)}_{t^*,(t^*-1)c}\right] \mid \text{Model } 0\right)$$

$$= \sum_c E \ldots E\left(\left[Y_{t_B c} LR^{(k)}_{t,(t-1)c} LR^{(k+1)}_{(t-1),(t-2)c} \prod_{t^*=t_B+1}^{t-2} LR^{(3)}_{t^*,(t^*-1)c}\right] \mid \text{Model } 0, Y_{(t_B+1)c}, \ldots, Y_{(t-1)c}\right)$$

$$= \sum_c E \ldots E\left(\left[Y_{t_B c} \rho_{tc} LR^{(k+1)}_{(t-1),(t-2)c} \prod_{t^*=t_B+1}^{t-2} LR^{(3)}_{t^*,(t^*-1)c}\right] \mid \text{Model } 0, Y_{(t_B+1)c}, \ldots, Y_{(t-2)c}\right)$$

$$= \ldots = \sum_c \left(Y_{t_B c} \prod_{t^*=t_B+1}^{t} \rho_{t^* c}\right) = E\left(Y_t \mid \text{Model } 0, \mathbf{Y}_{t_B c}\right)$$

An implicit assumption of the current weighted link relative estimator is that, within an estimation cell, establishments not reporting data for both months t and $t-1$ (which includes nonsampled units, and late reporting and nonresponse units in month t, as well as preliminary reporting units in month t for which data were not reported in month $t-1$) have the same expected growth rate as establishments reporting data, i.e., they all follow Model 0.

A more reasonable assumption may be that the proportionality factor varies not only by the static characteristics currently used to define estimation cells, but also by dynamic characteristics related to recent employment information. If, instead of Model 0, proportionality factors vary across classifications of establishments within estimation cell

Model 1: $E\left(Y_{tcgi} \mid Y_{(t-1)cgi}\right) = \rho_{tcg} Y_{(t-1)cgi}$

where g represents some classification of establishments within estimation cell c

ρ_{tcg} is the model parameter describing the month t expected growth rate for class g within cell c.

then the expected value under this model of the current weighted link relative no longer equals the expected value of the population total. This can be shown by first writing the deviation of the ρ_{tcg} from ρ_{tc} as

$$\rho_{tcg} = \rho_{tc} + \delta_{tcg}$$

The expected value of the current weighted link relative under Model 1 is then

$$E\left(LR_{t,(t-1)c}^{(k)}\} \mid \text{Model}\,1\right) = E\left[\left(\frac{\sum_{i \in s_{t,(t-1)c|k}} w_{ci} Y_{tci}}{\sum_{i \in s_{t,(t-1)c|k}} w_{ci} Y_{(t-1)ci}}\right) \mid \text{Model}\,1\right]$$

$$= E\left[\left(\frac{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{tcgi}}{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}\right) \mid \text{Model}\,1\right] = \frac{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}E\left(Y_{tcgi} \mid \text{Model}\,1\right)}{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}$$

$$= \frac{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}\rho_{tcg}Y_{(t-1)cgi}}{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}} = \frac{\sum_{g}\rho_{tcg}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}$$

$$= \frac{\sum_{g}\rho_{tc}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi} + \sum_{g}\delta_{tcg}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}$$

$$= \rho_{tc} + \frac{\sum_{g}\delta_{tcg}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}} = \rho_{tc} + \sum_{g}\delta_{tcg}\,\hat{p}^{(k)}_{(t-1)cg} = \rho_{tc} + \Psi^{(k)}_{tc}$$

where $\hat{p}^{(k)}_{(t-1)cg} = \frac{\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}{\sum_{g}\sum_{i\in s_{t,(t-1)cg|k}} w_{cgi}Y_{(t-1)cgi}}$ is an estimate of $p^{(k)}_{(t-1)cg} = \frac{\sum_{i\in c,g} w_{cgi}Y_{(t-1)cgi}}{\sum_{g}\sum_{i\in c,g} w_{cgi}Y_{(t-1)cgi}}$, the proportion of the population total for estimation cell c contained within class g as of revision k.

Thus, classifications of sample below the estimation cell level, g, that result in deviances from the growth rate at the estimation cell level, ρ_{tc}, indicate the potential for errors in the current weighted link relative estimator. To the extent the estimated relative sizes of these classes are such that the deviations do not net out (i.e., $\Psi^{(k)}_{tc} \neq 0$), the current weighted link relative estimator will be biased under Model 1. Empirical information on these components is provided in section B of this chapter.

Note that, under complete response, the design expectation for $\Psi_{tc}^{(k)} = 0$, and the weighted link relative is unbiased under Model 1.

The expected value of the estimated employment for month t under Model 1 is

$$E\left(\hat{Y}_t^{(k)} \mid \text{Model}\,1\right) = \sum_c E\left(\left[Y_{t_B c}\, LR_{t,(t-1)c}^{(k)}\, LR_{(t-1),(t-2)c}^{(k+1)} \prod_{t^*=t_B+1}^{t-2} LR_{t^*,(t^*-1)c}^{(3)}\right] \mid \text{Model}\,1\right)$$

$$= \sum_c E \ldots E\left(\left[Y_{t_B c}\, LR_{t,(t-1)c}^{(k)}\, LR_{(t-1),(t-2)c}^{(k+1)} \prod_{t^*=t_B+1}^{t-2} LR_{t^*,(t^*-1)c}^{(3)}\right] \mid \text{Model}\,1, Y_{(t_B+1)c}, \ldots, Y_{(t-1)c}\right)$$

$$= \sum_c E \ldots E\left(\left[Y_{t_B c}\left(\rho_{tc} + \Psi_{tc}^{(k)}\right) LR_{(t-1),(t-2)c}^{(k+1)} \prod_{t^*=t_B+1}^{t-2} LR_{t^*,(t^*-1)c}^{(3)}\right] \mid \text{Model}\,1, Y_{(t_B+1)c}, \ldots, Y_{(t-2)c}\right)$$

$$= \ldots = \sum_c \left(Y_{t_B c}\left(\rho_{tc} + \Psi_{tc}^{(k)}\right)\left(\rho_{(t-1)c} + \Psi_{(t-1)c}^{(k+1)}\right) \prod_{t^*=t_B+1}^{t-2}\left(\rho_{t^*c} + \Psi_{t^*c}^{(3)}\right)\rho_{t^*c}\right)$$

$$= \sum_c \left(Y_{t_B c}\left[\prod_{t^*=t_B+1}^{t} \rho_{t^*c} + \Psi_{tc}^{(k)} \prod_{t^*=t_B+1}^{t-1} \rho_{t^*c} + \Psi_{(t-1)c}^{(k+1)} \rho_{tc} \prod_{t^*=t_B+1}^{t-2} \rho_{t^*c} + \sum_{t^*=t_B+1}^{t-2} \Psi_{t^*c}^{(3)} \prod_{t' \neq t^*} \rho_{t'c} + \ldots\right.\right.$$

$$\ldots + \rho_{tc}\Psi_{(t-1)c}^{(k+1)} \prod_{t^*=t_B+1}^{t-2} \Psi_{t^*c}^{(3)} \rho_{t^*c} + \rho_{(t-1)c}\Psi_{tc}^{(k)} \prod_{t^*=t_B+1}^{t-2} \Psi_{t^*c}^{(3)} + \sum_{t^*=t_B+1}^{t-2} \rho_{t^*c}\Psi_{tc}^{(k)}\Psi_{(t-1)c}^{(k+1)} \prod_{t' \neq t^*} \Psi_{t'c}^{(3)} +$$

$$\left.\left.\Psi_{tc}^{(k)}\Psi_{(t-1)c}^{(k+1)} \prod_{t^*=t_B+1}^{t-2} \Psi_{t^*c}^{(3)}\right]\right)$$

$$= E\left(Y_t \mid \text{Model}\,0\right) + \sum_c \left(Y_{t_B c}\left[\Psi_{tc}^{(k)} \prod_{t^*=t_B+1}^{t-1} \rho_{t^*c} + \Psi_{(t-1)c}^{(k+1)} \rho_{tc} \prod_{t^*=t_B+1}^{t-2} \rho_{t^*c} + \sum_{t^*=t_B+1}^{t-2} \Psi_{t^*c}^{(3)} \prod_{t' \neq t^*} \rho_{t'c} + \ldots\right.\right.$$

$$\ldots + \rho_{tc}\Psi_{(t-1)c}^{(k+1)} \prod_{t^*=t_B+1}^{t-2} \Psi_{t^*c}^{(3)} \rho_{t^*c} + \rho_{(t-1)c}\Psi_{tc}^{(k)} \prod_{t^*=t_B+1}^{t-2} \Psi_{t^*c}^{(3)} + \sum_{t^*=t_B+1}^{t-2} \rho_{t^*c}\Psi_{tc}^{(k)}\Psi_{(t-1)c}^{(k+1)} \prod_{t' \neq t^*} \Psi_{t'c}^{(3)} +$$

$$\left.\left.\Psi_{tc}^{(k)}\Psi_{(t-1)c}^{(k+1)} \prod_{t^*=t_B+1}^{t-2} \Psi_{t^*c}^{(3)}\right]\right)$$

This calculation assumes the number of sample units in $s_{t,(t-1)c|k}$ is sufficiently large so that the expectation of the product of ratios is approximately equal to the product of the expectations of the ratios. Again, under complete response, the design expectation for $\Psi_{tc}^{(k)} = 0$, and the estimated employment is unbiased under Model 1.

Assuming ρ_{t^*c} and $\Psi_{t^*c}^{(k)}$ are relatively stable across time, replacing with mean values, $\overline{\rho}_c$ and $\overline{\Psi}_c$ yields

$$E\left(\hat{Y}_t \mid \text{Model 1}\right) = E\left(Y_t \mid \text{Model 0}\right) +$$

$$\sum_c \left[Y_{t_Bc} \left(\sum_{t^*=t_B+1}^{t} \overline{\Psi}_c \overline{\rho}_c^{(t-t_B-1)} + \ldots + \sum_{t^*=t_Bc+1}^{t} \overline{\Psi}_c^{(t-t_B-1)} \overline{\rho}_c + \overline{\Psi}_c^{(t-t_B)} \right) \right]$$

$$= E\left(Y_t \mid \text{Model 0}\right) +$$

$$\sum_c \left[Y_{t_Bc} \left(\left(t - t_B\right) \overline{\Psi}_c \overline{\rho}_c^{(t-t_B-1)} + \ldots + \left(t - t_B\right) \overline{\Psi}_c^{(t-t_B-1)} \overline{\rho}_c + \overline{\Psi}_c^{(t-t_B)} \right) \right]$$

Further, assuming $\overline{\Psi}_c$ is small relative to $\overline{\rho}_c$ (if $\overline{\rho}_c$ is around 1.0, say $\overline{\Psi}_c$ <0.001), then 2nd and higher order terms including $\overline{\Psi}_c$ may reasonably be ignored, leaving

$$E\left(\hat{Y}_t \mid \text{Model 1}\right) = E\left(Y_t \mid \text{Model 0}\right) + \sum_c \left[Y_{t_Bc} \left(\left(t - t_B\right) \overline{\Psi}_c \overline{\rho}_c^{(t-t_B-1)} \right) \right]$$

This result shows the bias in $\hat{Y}_t$ due to model misspecification increases with the number of months from the last benchmark date, assuming $\overline{\Psi}_c$ is non-zero. This provides the motivation for carrying out benchmark updates on a frequent basis. For the CES survey, the number of months from the last benchmark ranges from 11 to 23. Thus, even if biases on the monthly link relatives are less than 0.001, the bias on the monthly employment estimate could be on the order of one percent of the population value.

Given incomplete reporting, the expected value of the weighted link relative under Model 1 will vary between the preliminary and the final due to the inclusion of late reporters. The expected difference can be written as

$$E\left(LR_{t,(t-1)c}^{(2)} - LR_{t,(t-1)c}^{(0)}\right) = \left(\rho_{tc} + \sum_g \delta_{tcg}\,\hat{p}_{tcg}^{(2)}\right) - \left(\rho_{tc} + \sum_g \delta_{tcg}\,\hat{p}_{tcg}^{(0)}\right)$$

$$= \sum_g \delta_{tcg}\,\hat{p}_{tcg}^{(2)} - \sum_g \delta_{tcg}\,\hat{p}_{tcg}^{(0)} = \sum_g \delta_{tcg}\left(\hat{p}_{tcg}^{(2)} - \hat{p}_{tcg}^{(0)}\right)$$

To the extent the estimated relative sizes of the population in estimation cell c contained within class g vary between preliminary and final, the preliminary and final link relatives will differ. Empirical information on these values is provided in section B of this chapter. One approach to generation of a preliminary estimate subject to less revision would be to utilize the sample that can later be included as late reporters, thereby reducing differences between the $\hat{p}_{(t-1)cg}^{(1)}$ and $\hat{p}_{(t-1)cg}^{(3)}$. This is the approach developed in the remainder of this chapter.

B. Potential for Error in Current CES Estimation Methodology

1. Indirect Indicators of Error Due to Late Reporting

Commonly, indirect indicators of the impact of nonreporting are used to assess the potential impact, as data for the nonreporters are not known. The CES survey provides more tangible information related to the impact of nonreporting through 2nd and 3rd closing revisions (late reporting) and, to a lesser extent, benchmark revisions (nonresponse, plus sampling and measurement error).

Comparisons of first and third closing estimates provide a direct indication of the impact of late reporting, as the only difference between the two estimates is the inclusion of late reporters into the sample. The relative difference between first and third closing estimates for month t,

$$\text{RelDiff}_{t(0,2)} = \frac{\hat{Y}_t^{(0)} - \hat{Y}_t^{(2)}}{\hat{Y}_t^{(2)}} \times 100\%$$

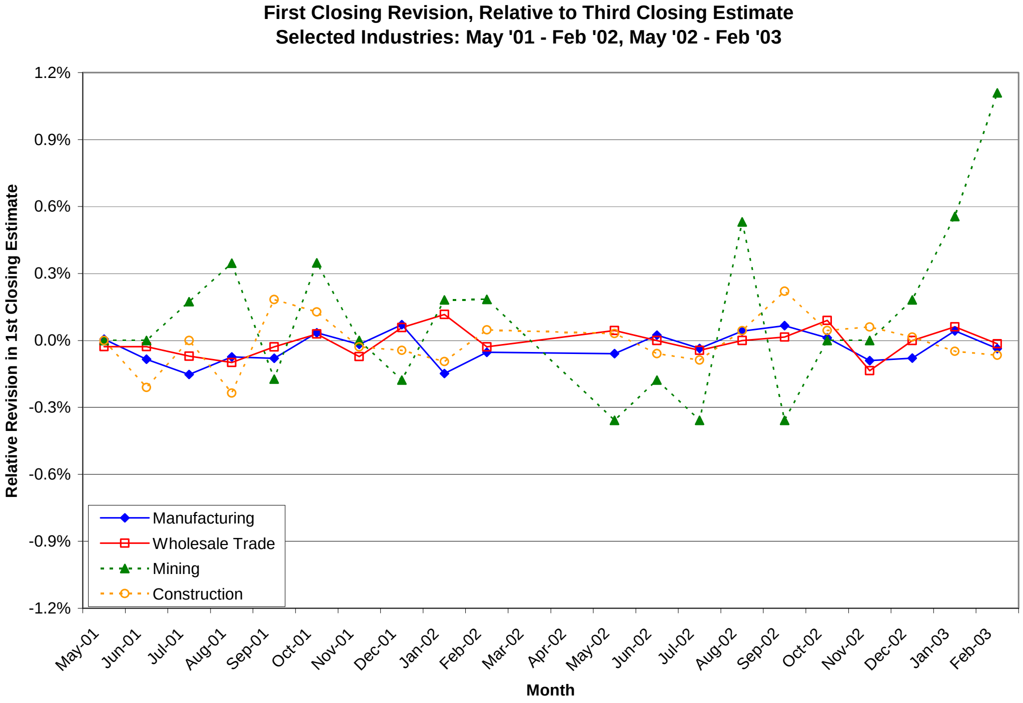
and the difference between first and third closing estimates of the month-to-month change from month $t-1$ to month t,

$$\text{Diff}_{t(0,2)} = \left[\hat{Y}_t^{(0)} - \hat{Y}_{(t-1)}^{(1)}\right] - \left[\hat{Y}_t^{(2)} - \hat{Y}_{(t-1)}^{(2)}\right] = \Delta_{t,(t-1)}^{(0)} - \Delta_{t,(t-1)}^{(2)}$$

provide measures of the extent to which growth rates for late reporters differed from those for early reporters. Large differences provide an indication that the late reporting mechanism may not be ignorable.

Figure 24 shows relative differences between first and third closing published non-seasonally adjusted estimates of monthly employment for the period May 2001 – February 2002 and May 2002 – February 2003. March and April were excluded from this graph due to the nature of CES survey processing as, for these years, annual benchmark data were incorporated with the publication of first closing estimates for May (and thus second closing for April and third closing for March) thereby negating the ability to measure solely late reporting impact for these months. Although the larger industries have experienced fairly small revisions (absolute relative differences less than 0.3%), the revisions for Mining have been much greater, with the absolute relative difference as high as 1.1% in February 2003.

Figure 24-First Closing Revision



Revisions in the monthly employment estimates and in the estimates of month-to-month change in employment can also be compared with the month-to-month change in employment, which is a primary measure for assessing the employment data. Revisions that are large relative to the estimated change could serve to decrease the utility of the preliminary reports. Magnitudes of the revisions in monthly and month-to-month change in employment to the first closing estimate of month-to-month change in employment for the period May 2002 – February 2003 are provided in Table 32.

Table 32-First Closing Revision versus Month-to-Month Change

First Closing Revisions versus First Closing Month-to-Month Employment Change
Selected Industries: May '02 - Feb '03
(Numbers in thousands)

	Manufacturing				Wholesale Trade				Mining				Construction			
	Employment (1st Closing)	Revision in 1st Closing Employment	Employment Change from Prior Month (1st Closing)	Revision in 1st Closing Employment Change	Employment (1st Closing)	Revision in 1st Closing Employment	Employment Change from Prior Month (1st Closing)	Revision in 1st Closing Employment Change	Employment (1st Closing)	Revision in 1st Closing Employment	Employment Change from Prior Month (1st Closing)	Revision in 1st Closing Employment Change	Employment (1st Closing)	Revision in 1st Closing Employment	Employment Change from Prior Month (1st Closing)	Revision in 1st Closing Employment Change
May-02	16,769	-10	24	-9	6,682	3	19	4	561	-2	4	-2	6,595	2	196	1
Jun-02	16,838	4	78	5	6,713	0	27	1	562	-1	3	-1	6,794	-4	200	-7
Jul-02	16,755	-6	-88	-5	6,716	-3	3	-3	561	-2	-1	-1	6,857	-6	61	0
Aug-02	16,784	7	30	12	6,698	0	-15	0	562	3	3	3	6,864	3	13	3
Sep-02	16,709	11	-70	-1	6,672	1	-27	2	561	-2	-4	-2	6,785	15	-78	11
Oct-02	16,643	2	-74	-1	6,667	6	-6	6	560	0	0	1	6,752	3	-50	5
Nov-02	16,575	-15	-73	-12	6,662	-9	-11	-9	554	0	-6	0	6,645	4	-111	5
Dec-02	16,487	-13	-72	-14	6,646	0	-7	0	550	1	-4	1	6,448	1	-196	-4
Jan-03	16,341	7	-136	10	6,585	4	-62	5	537	3	-14	3	6,128	-3	-323	-1
Feb-03	16,293	-6	-58	-3	6,584	-1	-4	-2	535	6	-3	4	6,065	-4	-66	2

Although revisions for several months are larger than the first closing estimate of month-to-month employment change, the changes in these situations are small. For months with larger employment changes, revisions are not of the magnitude of the change, but could nonetheless be viewed as substantial (five of eighteen first closing changes of at least 50,000 saw a revision in the first closing estimated employment level that was 10%+ of the magnitude of the first closing estimated change (i.e., $\left|Y_t^{(2)} - Y_t^{(0)}\right| > 0.1 * \Delta_{t,(t-1)}^{(0)}$), while four saw a 10%+ revision in the magnitude of the change (i.e., $\left|\Delta_t^{(2)} - \Delta_t^{(0)}\right| > 0.1 * \Delta_{t,(t-1)}^{(0)}$)). Viewed from this perspective, late reporting could be considered to have an adverse impact on the accuracy of the first closing estimates.

2. Differences between Preliminary and Late Reporters

As discussed in section A of this chapter, to the extent there is misspecification in the underlying model upon which the current CES weighted link relative is based (Model 0), there is the potential for error in the resulting estimated employment. Of particular interest for this research is that model misspecification could result in differences between preliminary and revised estimates.

One way this potential for error due to model misspecification can be assessed is to look at the level of agreement in weighted link relatives between preliminary and late reporters. If Model 0 fits well, then over time the relationship between weighted link relatives for preliminary and late reporters within an estimation cell should follow a straight line through the origin with a slope of 1. Figure 25-Figure 28 show the actual preliminary and late reporter weighted link relatives for March 2000 through December 2002. The straight line assuming Model 0 is provided, along with the straight line fitted to the data points. As can be seen, for both Construction (slope = 0.963) and Manufacturing (slope = 0.905) the fitted line has a slope close to 1. This is not the case for Mining (slope = 0.411) and Wholesale Trade (slope = 0.671). Note the scales on each figure are different, to allow better visibility to the data points for that industry.

Figure 25-Link Relatives for Preliminary, Late Reporters-Construction

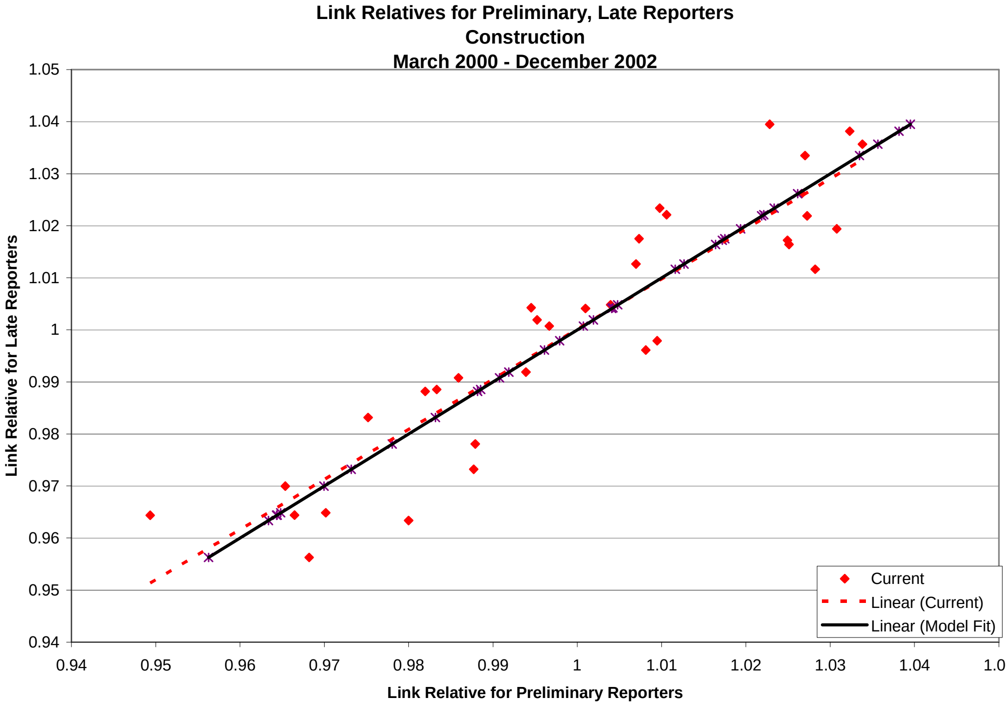


Figure 26-Link Relatives for Preliminary, Late Reporters-Manufacturing

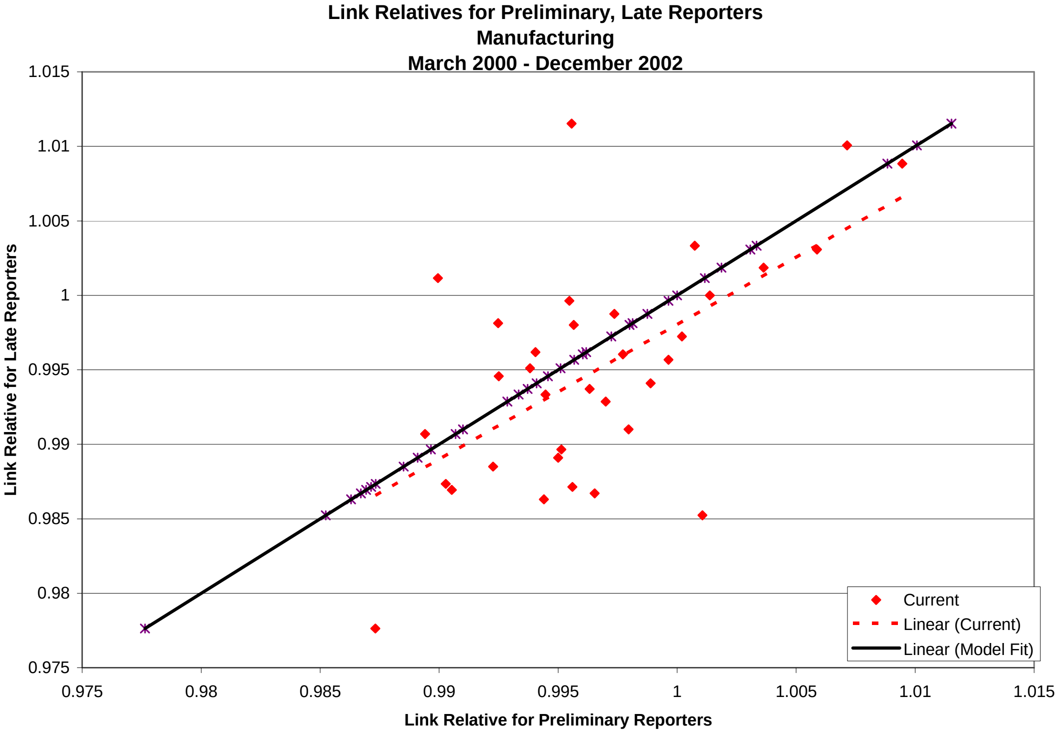


Figure 27-Link Relatives for Preliminary, Late Reporters-Mining

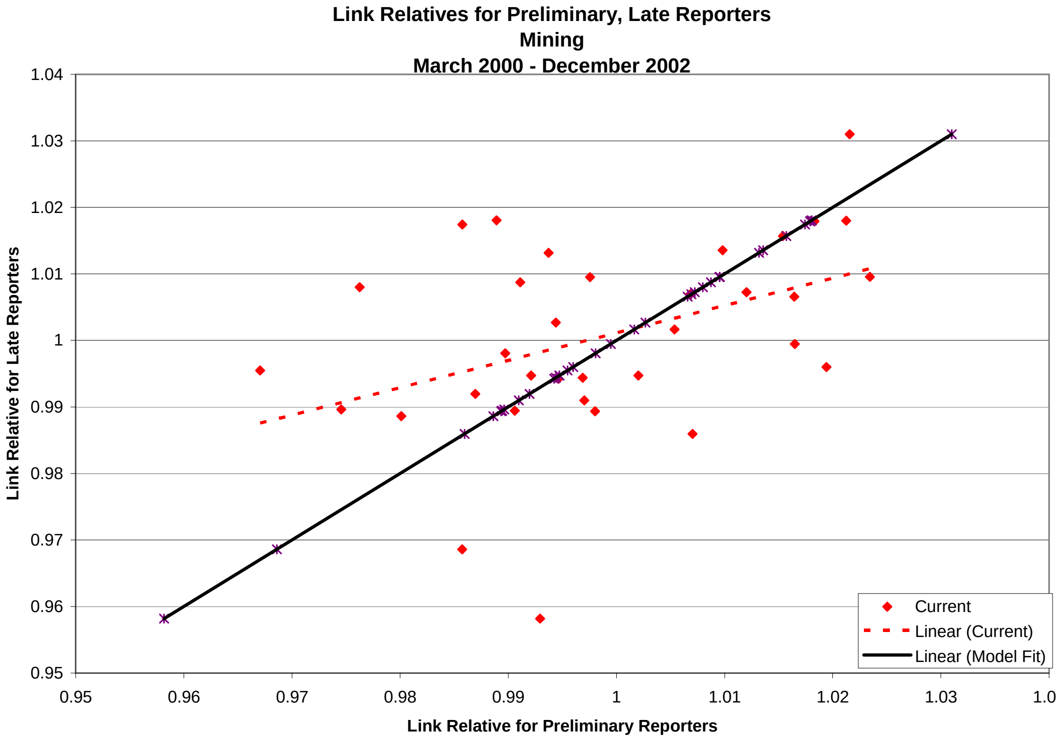
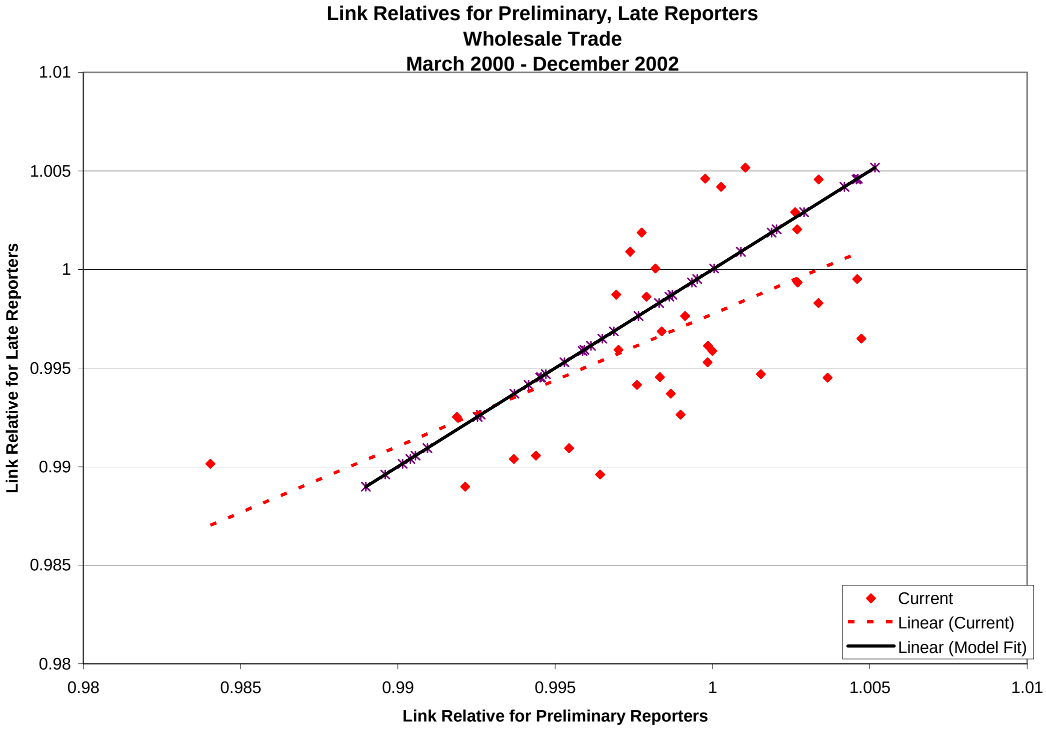
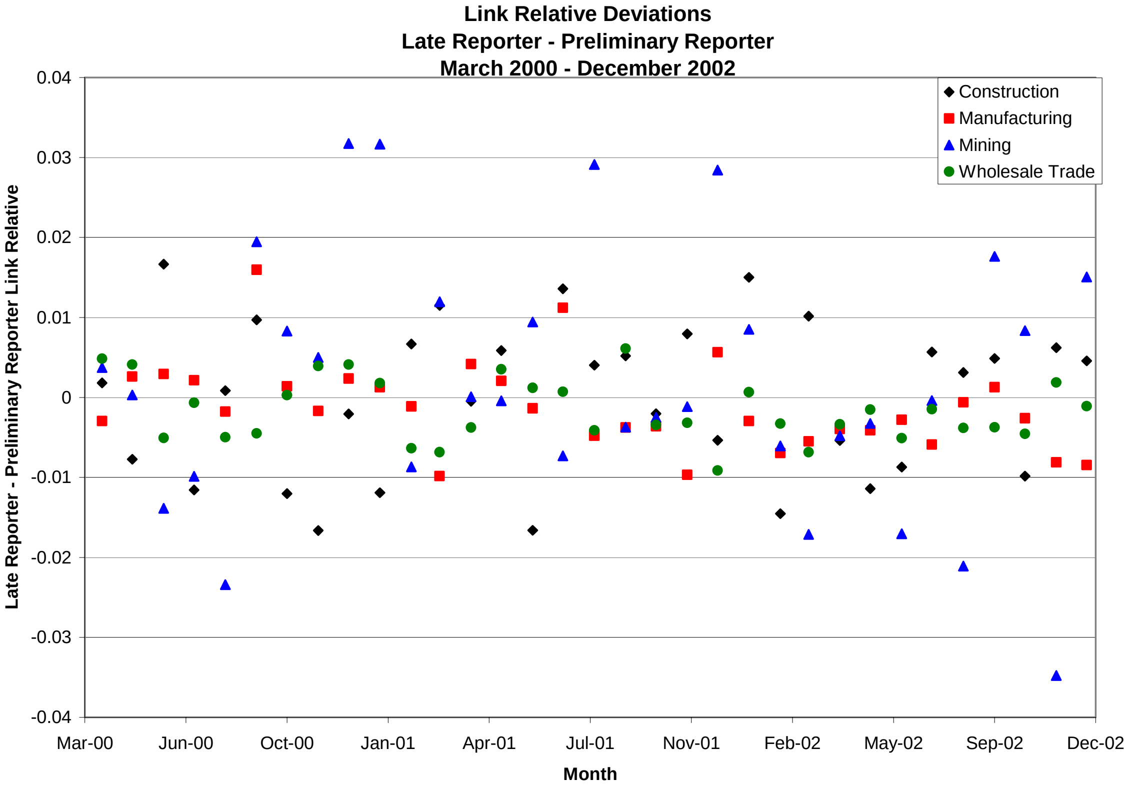


Figure 28-Link Relatives for Preliminary, Late Reporters-Wholesale Trade



Regardless of fit, weighted link relatives for late reporters occasionally differ from those of preliminary reporters by more than one percentage point, as illustrated in Figure 29. It is these more extreme deviations that will tend to yield larger revisions, and which the approach developed in the next section is intended to control for. In order to develop the approach, a set of underlying factors that may be driving these deviations must be identified.

Figure 29-Link Relative Deviations: Late Reporters – Preliminary Reporters



3. Components of Model Misspecification Error

Potential model misspecification may be more directly assessed by examining the components of error defined in section A of this chapter: δ_{tcg} (deviation of class growth rate from cell growth rate); and $\hat{p}^{(2)}_{(t-1)cg}$ (estimated proportion of cell contained within class), identified in the previous section. The question is what characteristics should be used to define classes within estimation cell.

Two sets of characteristics were hypothesized to be related to employment growth rate for month t: prior month employment size and prior month employment change. Employment size was considered because: 1) growth rate experience may reasonably be expected to differ for small and large establishments; and 2) growth rates are inherently more unstable for establishments with smaller employment in month $t-1$ (i.e., an employment change of 1 for an establishment of with month $t-1$ employment of 5 represents a 20% change). Prior month employment change was considered as employment change for the current period could vary based upon the relative size of the employment change for the immediately prior period.

Employment change can be viewed in actual $\left(Y_{(t-1)i} - Y_{(t-2)i}\right)$ or relative $\left(Y_{(t-1)i} / Y_{(t-2)i}\right)$ terms. For smaller establishments, actual employment change provides a more stable measure than does relative employment change, while the opposite is true for larger establishments. Therefore, the approach was developed to use actual employment change for smaller establishments and relative employment change for larger establishments.

Rank ordered prior month employment changes (both actual and relative) for each month were separated into three sets of units for purposes of defining prior month employment change classes within an industry. Establishments within the first set were designated as low prior month employment change, those within the second set were designated as mid prior month employment change and those within the third set were designated as high prior month employment change. Those units for which prior employment change was not known (i.e., unit did not report for month $t-2$) were designated as unknown prior month employment change.

The class utilized for an establishment was determined based upon the establishment’s employment level for month $t-1$ (the base month for the employment change to be estimated by the model). For establishments classified as small employment level (<50) for month $t-1$, the actual prior month employment change class was used; establishments classified as large employment level for month $t-1$, the relative prior month employment change class was used.

Average values for δ_{tcg} and $\hat{p}^{(2)}_{(t-1)cg}$ for the period March 2000 – December 2002, based upon design size class within industry, were calculated using the final reported sample (i.e., preliminary plus late reporters). For δ_{tcg}, the standard deviation of the monthly values was also calculated, along with the number of monthly values that were greater than zero (to provide an indication of consistency of direction). For $\hat{p}^{(2)}_{(t-1)cg}$, the minimum and maximum monthly values were calculated to indicate the range for possible use in estimating potential error associated with the current weighted link relative. Average numbers of total and preliminary reporters were also calculated as an indication of whether sufficient sample sizes exist for estimation of parameters.

Looking at prior month size class (Table 33), it appears that the smallest establishments (<10) have different employment growth rates (δ_{tcg} ranges from 0.006 to 0.021 and, with the exception of Mining, values of δ_{tcg} were positive for 85%+ of the months).

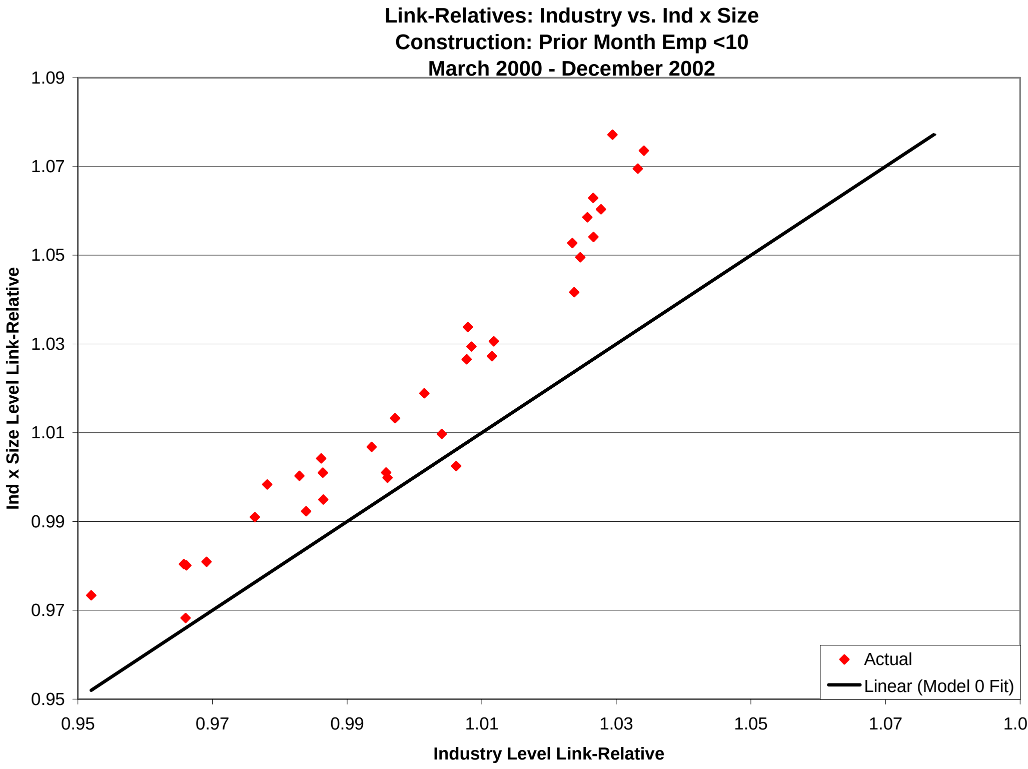
Table 33-Components of Model Misspecification Error: Size

Components of Model Misspecification Error
by Prior Month Employment Size
March 2000 - December 2002

Industry	Prior Month Employment Size	Average Sample Size		Values for δ_{tcg}			Values for $\hat{p}_{(t-1)cg}$		
		Total	Preliminary Reporters	Average	stdev	Percent >0	Average	Min	Max
Construction	<10	2692	2344	0.0191	0.0114	97.1%	0.1678	0.1590	0.1835
	10-19	1239	1086	-0.0015	0.0072	47.1%	0.1286	0.1122	0.1441
	20-49	1804	1571	-0.0018	0.0066	41.2%	0.2077	0.1823	0.2212
	50-99	1473	1271	-0.0039	0.0066	20.6%	0.1573	0.1440	0.1656
	100-249	1435	1230	-0.0050	0.0089	23.5%	0.1902	0.1693	0.2196
	250+	465	382	-0.0075	0.0184	32.4%	0.1484	0.1200	0.1722
Manufacturing	<10	1554	1277	0.0128	0.0083	100.0%	0.0232	0.0193	0.0267
	10-19	1231	1043	0.0021	0.0061	61.8%	0.0308	0.0264	0.0359
	20-49	2323	1983	0.0017	0.0054	61.8%	0.0730	0.0653	0.0835
	50-99	3182	2660	-0.0001	0.0042	44.1%	0.0972	0.0891	0.1054
	100-249	6180	5141	0.0009	0.0050	55.9%	0.2155	0.2029	0.2265
	250+	4687	3735	-0.0012	0.0021	29.4%	0.5604	0.5379	0.5835
Mining	<10	285	233	0.0212	0.0348	67.6%	0.0711	0.0557	0.0838
	10-19	214	175	-0.0166	0.0345	32.4%	0.0680	0.0575	0.0805
	20-49	317	252	-0.0034	0.0197	47.1%	0.1333	0.1111	0.1563
	50-99	191	151	-0.0006	0.0101	41.2%	0.0953	0.0787	0.1130
	100-249	163	125	0.0094	0.0255	58.8%	0.1326	0.0978	0.1737
	250+	122	85	-0.0022	0.0097	38.2%	0.4997	0.4494	0.5519
Wholesale Trade	<10	2252	1753	0.0057	0.0052	85.3%	0.1385	0.0527	0.1542
	10-19	1009	801	0.0005	0.0090	50.0%	0.1079	0.0737	0.1191
	20-49	1112	851	-0.0007	0.0055	47.1%	0.1756	0.1565	0.1946
	50-99	706	530	-0.0019	0.0082	50.0%	0.1474	0.1330	0.1666
	100-249	891	660	-0.0006	0.0044	50.0%	0.1916	0.1676	0.2302
	250+	423	293	-0.0013	0.0047	26.5%	0.2389	0.1851	0.3202

This situation is illustrated in Figure 30, which graphs weighted link relatives for reporting establishments in the prior month employment class <10 against the weighted link relatives for Construction as a whole. In this case, the link relatives for the industry as a whole are almost consistently below those for this size class. These results suggest the use of prior month employment size, perhaps collapsed into two or a few classes, in Model 1 could better explain employment growth rates for potential late reporters than Model 0.

Figure 30- Comparison of Industry and Industry x Size Link Relatives



The results contained in Table 34 provide the breakout of values for δ_{tcg} and $\hat{p}^{(2)}_{(t-1)cg}$ for prior month employment size class by prior month employment change class. These results indicate that for smaller establishments, particularly those in the <10 size class, prior month employment change that was low or high deviate noticeably from the industry level growth rate, and in opposite directions. Establishments with prior month employment of 10-19 and 20-49 showed some tendencies in this same direction, but not to the extent seen for the smallest size class. Where deviations occurred, establishments with low prior month employment change experienced growth rates larger than those for the industry as a whole, while establishments with high prior month employment change experienced growth rates smaller than those for the industry as a whole.

Table 34-Components of Model Misspecification Error: Size x Change

Components of Model Misspecification Error: Construction
by Prior Month Employment Size x Prior Month Employment Change
March 2000 - December 2002

Industry	Prior Month Employment Size	Prior Month Employment Change	Average Sample Size		Values for δ_{tcg}			Values for $\hat{p}_{(t-1)cg}$		
			Total	Preliminary Reporters	Average	stdev	Percent >0	Average	Min	Max
Construction	<10	Low	492	424	0.1054	0.0396	100.0%	0.0278	0.0202	0.0349
		Mid	1708	1501	0.0105	0.0119	82.4%	0.1021	0.0931	0.1184
		High	374	322	-0.0307	0.0208	11.8%	0.0310	0.0221	0.0389
		Unk	118	97	0.0286	0.0445	73.5%	0.0070	0.0027	0.0134
	10-19	Low	383	335	0.0279	0.0175	97.1%	0.0347	0.0257	0.0430
		Mid	443	391	-0.0061	0.0100	20.6%	0.0493	0.0406	0.0767
		High	378	330	-0.0209	0.0175	8.8%	0.0405	0.0299	0.0538
		Unk	36	30	-0.0044	0.0613	38.2%	0.0040	0.0019	0.0079
	20-49	Low	571	498	0.0031	0.0167	58.8%	0.0575	0.0448	0.0736
		Mid	592	516	-0.0075	0.0097	26.5%	0.0729	0.0570	0.0946
		High	597	522	0.0008	0.0120	50.0%	0.0711	0.0518	0.0835
		Unk	44	34	-0.0123	0.0507	47.1%	0.0061	0.0020	0.0126
	50-99	Low	459	397	-0.0139	0.0239	32.4%	0.0437	0.0380	0.0517
		Mid	484	417	-0.0050	0.0112	41.2%	0.0540	0.0491	0.0608
		High	499	432	0.0052	0.0129	67.6%	0.0553	0.0475	0.0630
		Unk	31	24	-0.0016	0.0408	52.9%	0.0044	0.0018	0.0101
	100-249	Low	418	360	-0.0202	0.0197	17.6%	0.0525	0.0394	0.0648
		Mid	456	390	-0.0064	0.0151	38.2%	0.0604	0.0499	0.0735
		High	532	458	0.0066	0.0128	70.6%	0.0721	0.0613	0.0825
		Unk	29	22	0.0010	0.0476	38.2%	0.0051	0.0020	0.0125
	250+	Low	119	97	-0.0156	0.0410	23.5%	0.0359	0.0229	0.0551
		Mid	164	137	-0.0040	0.0154	44.1%	0.0542	0.0296	0.0784
		High	166	136	-0.0030	0.0376	47.1%	0.0518	0.0361	0.0765
		Unk	16	11	-0.0189	0.0362	23.5%	0.0065	0.0024	0.0279

Components of Model Misspecification Error: Manufacturing
by Prior Month Employment Size x Prior Month Employment Change
March 2000 - December 2002

Industry	Prior Month Employment Size	Prior Month Employment Change	Average Sample Size		Values for δ_{tcg}			Values for $\hat{p}_{(t-1)cg}$		
			Total	Preliminary Reporters	Average	stdev	Percent >0	Average	Min	Max
Manufacturing	<10	Low	260	202	0.0926	0.0401	100.0%	0.0033	0.0023	0.0043
		Mid	1071	899	0.0040	0.0090	67.6%	0.0159	0.0119	0.0194
		High	159	126	-0.0355	0.0218	2.9%	0.0030	0.0022	0.0039
		Unk	65	50	0.0355	0.0480	88.2%	0.0010	0.0003	0.0022
	10-19	Low	319	270	0.0237	0.0129	100.0%	0.0074	0.0055	0.0092
		Mid	604	517	-0.0021	0.0050	29.4%	0.0153	0.0118	0.0196
		High	271	227	-0.0119	0.0139	11.8%	0.0072	0.0052	0.0089
		Unk	38	29	0.0050	0.0205	64.7%	0.0009	0.0002	0.0019
	20-49	Low	795	680	0.0109	0.0143	82.4%	0.0234	0.0195	0.0287
		Mid	783	669	-0.0026	0.0055	29.4%	0.0245	0.0188	0.0321
		High	682	585	-0.0039	0.0089	35.3%	0.0231	0.0175	0.0302
		Unk	63	50	0.0072	0.0230	52.9%	0.0020	0.0008	0.0044
	50-99	Low	1081	908	0.0004	0.0106	52.9%	0.0304	0.0260	0.0353
		Mid	1042	869	-0.0025	0.0064	29.4%	0.0317	0.0276	0.0386
		High	991	833	0.0016	0.0077	61.8%	0.0326	0.0278	0.0377
		Unk	68	51	0.0053	0.0211	61.8%	0.0025	0.0011	0.0055
	100-249	Low	2076	1734	0.0030	0.0174	35.3%	0.0718	0.0650	0.0808
		Mid	1803	1496	-0.0022	0.0049	29.4%	0.0623	0.0541	0.0694
		High	2189	1830	0.0016	0.0054	76.5%	0.0770	0.0714	0.0861
		Unk	112	81	-0.0063	0.0292	52.9%	0.0043	0.0018	0.0082
	250+	Low	1326	1065	-0.0061	0.0081	17.6%	0.1436	0.1199	0.1865
		Mid	1759	1398	-0.0011	0.0038	38.2%	0.2294	0.1841	0.2663
		High	1476	1185	0.0035	0.0049	79.4%	0.1686	0.1408	0.1935
		Unk	126	88	-0.0048	0.0184	44.1%	0.0187	0.0069	0.0400

Components of Model Misspecification Error: Mining
by Prior Month Employment Size x Prior Month Employment Change
March 2000 - December 2002

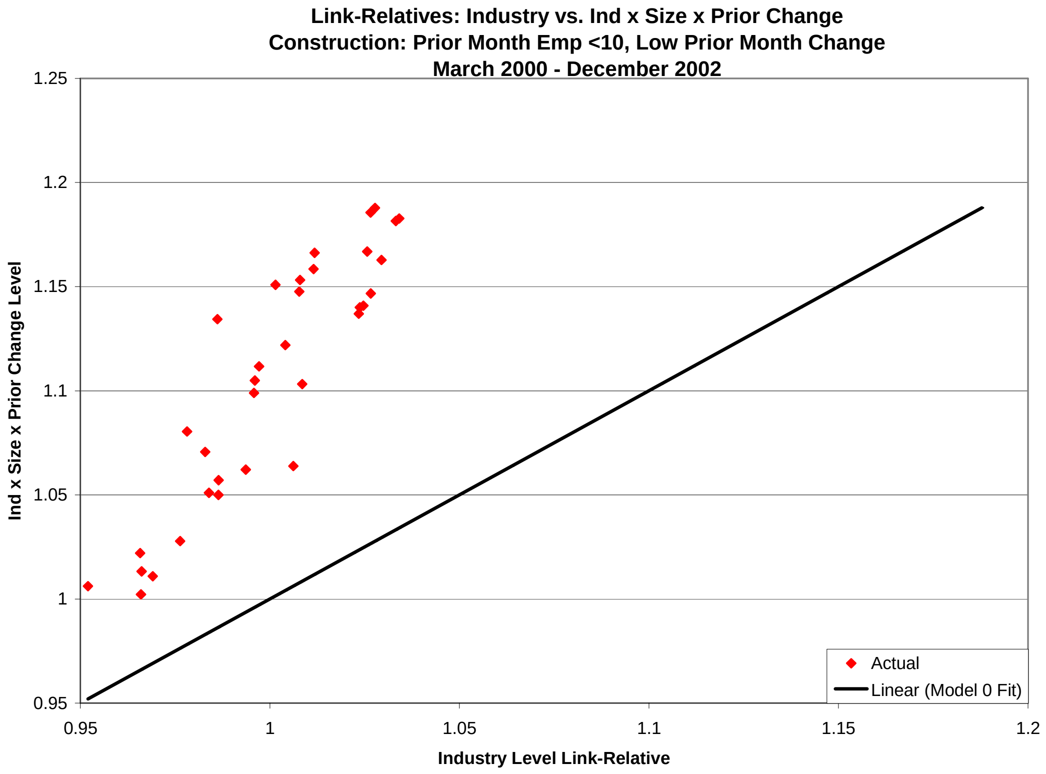
Industry	Prior Month Employment Size	Prior Month Employment Change	Average Sample Size		Values for δ_{tcg}			Values for $\hat{p}_{(t-1)cg}$		
			Total	Preliminary Reporters	Average	stdev	Percent >0	Average	Min	Max
Mining	<10	Low	45	36	0.1826	0.2594	85.3%	0.0086	0.0032	0.0164
		Mid	201	166	0.0010	0.0271	55.9%	0.0508	0.0369	0.0596
		High	30	24	0.0017	0.1279	47.1%	0.0089	0.0050	0.0167
		Unk	9	7	0.0761	0.4931	55.9%	0.0027	0.0000	0.0100
	10-19	Low	54	45	0.0169	0.0964	58.8%	0.0160	0.0077	0.0290
		Mid	104	84	-0.0172	0.0366	35.3%	0.0327	0.0221	0.0422
		High	51	42	-0.0442	0.0720	23.5%	0.0176	0.0077	0.0273
		Unk	5	4	-0.0183	0.1753	47.1%	0.0017	0.0000	0.0064
	20-49	Low	88	71	0.0079	0.0464	70.6%	0.0340	0.0157	0.0671
		Mid	123	98	-0.0051	0.0238	44.1%	0.0544	0.0284	0.0775
		High	97	78	-0.0169	0.0340	29.4%	0.0419	0.0182	0.0623
		Unk	8	6	0.0352	0.2509	55.9%	0.0030	0.0001	0.0117
	50-99	Low	56	45	0.0025	0.0278	55.9%	0.0269	0.0191	0.0400
		Mid	68	53	-0.0046	0.0149	29.4%	0.0324	0.0199	0.0480
		High	63	50	0.0022	0.0207	50.0%	0.0342	0.0169	0.0468
		Unk	4	3	-0.0371	0.1768	32.4%	0.0018	0.0000	0.0093
	100-249	Low	54	41	0.0050	0.0503	52.9%	0.0463	0.0254	0.0844
		Mid	48	37	-0.0037	0.0174	44.1%	0.0350	0.0199	0.0570
		High	58	44	0.0145	0.0395	67.6%	0.0487	0.0302	0.0778
		Unk	4	3	0.0156	0.0803	58.8%	0.0027	0.0003	0.0124
	250+	Low	34	23	-0.0083	0.0331	32.4%	0.1473	0.0435	0.3023
		Mid	46	32	0.0026	0.0126	55.9%	0.1913	0.0892	0.3479
		High	38	27	-0.0057	0.0223	38.2%	0.1473	0.0534	0.2482
		Unk	4	3	-0.0924	0.3837	38.2%	0.0138	0.0000	0.1282

Components of Model Misspecification Error: Wholesale Trade
by Prior Month Employment Size x Prior Month Employment Change
March 2000 - December 2002

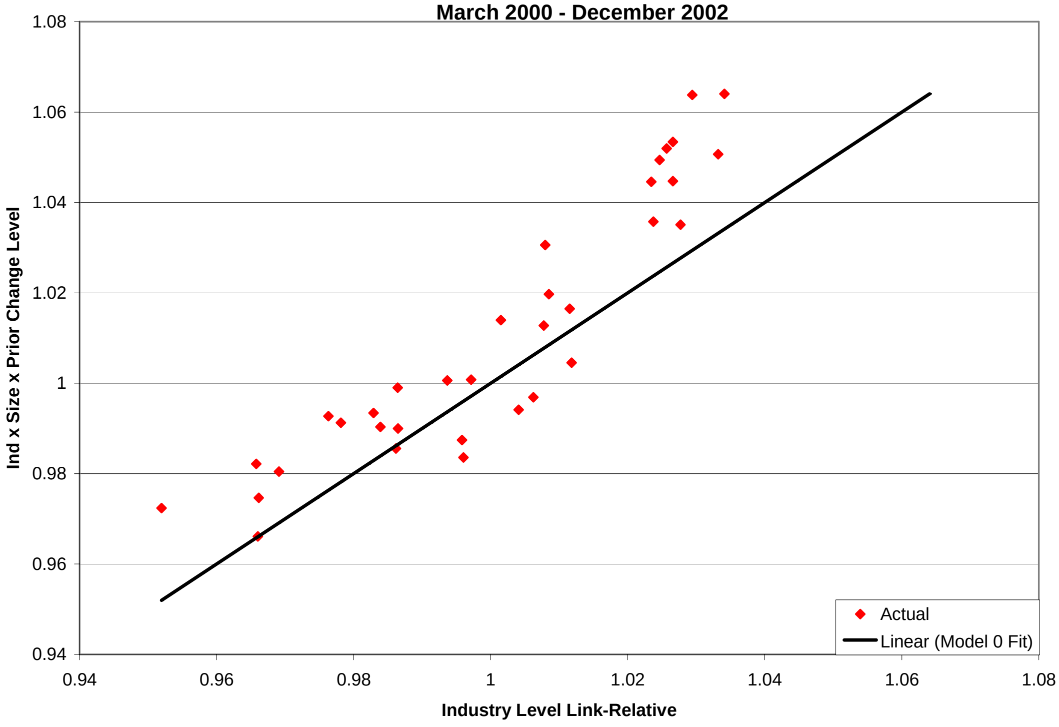
Industry	Prior Month Employment Size	Prior Month Employment Change	Average Sample Size		Values for δ_{tcg}			Values for $\hat{p}_{(t-1)cg}$		
			Total	Preliminary Reporters	Average	stdev	Percent >0	Average	Min	Max
Wholesale Trade	<10	Low	212	164	0.0564	0.0320	97.1%	0.0128	0.0050	0.0165
		Mid	1806	1419	0.0024	0.0067	61.8%	0.1078	0.0379	0.1209
		High	146	116	-0.0242	0.0315	17.6%	0.0123	0.0068	0.0160
		Unk	87	54	0.0188	0.0523	67.6%	0.0056	0.0008	0.0161
	10-19	Low	205	160	0.0173	0.0157	91.2%	0.0201	0.0145	0.0264
		Mid	584	470	-0.0011	0.0048	35.3%	0.0624	0.0402	0.0725
		High	187	151	-0.0147	0.0188	8.8%	0.0214	0.0132	0.0272
		Unk	33	20	0.0181	0.1031	61.8%	0.0039	0.0011	0.0109
	20-49	Low	328	247	0.0096	0.0168	82.4%	0.0476	0.0381	0.0647
		Mid	441	341	-0.0029	0.0050	29.4%	0.0712	0.0561	0.0872
		High	305	239	-0.0066	0.0091	23.5%	0.0494	0.0346	0.0630
		Unk	38	24	-0.0054	0.0205	44.1%	0.0075	0.0007	0.0197
	50-99	Low	226	169	0.0010	0.0172	70.6%	0.0428	0.0324	0.0526
		Mid	232	175	-0.0009	0.0058	44.1%	0.0494	0.0385	0.0652
		High	226	173	-0.0045	0.0153	41.2%	0.0501	0.0420	0.0668
		Unk	23	14	-0.0031	0.0797	38.2%	0.0051	0.0008	0.0118
	100-249	Low	286	213	0.0009	0.0085	64.7%	0.0584	0.0414	0.0744
		Mid	272	200	0.0006	0.0079	50.0%	0.0597	0.0392	0.0943
		High	307	231	-0.0015	0.0052	38.2%	0.0656	0.0529	0.0982
		Unk	26	15	-0.0124	0.0471	41.2%	0.0079	0.0018	0.0179
	250+	Low	118	81	-0.0051	0.0135	35.3%	0.0696	0.0367	0.1288
		Mid	154	106	0.0004	0.0064	47.1%	0.0786	0.0511	0.1140
		High	134	96	0.0017	0.0066	55.9%	0.0693	0.0488	0.1193
		Unk	18	10	-0.0026	0.0296	50.0%	0.0214	0.0019	0.0710

Again, several illustrations show the degree of deviation from Model 0 for selected classes. Figure 31 presents graphs of weighted link relatives for reporting establishments in the prior month employment class <10 for Construction, by prior month employment change class (Low, Mid, High), against the weighted link relatives for the industry as a whole. If Model 0 fit for all classes, the observations would be on the 45 degree line denoted as "Linear (Model 0 Fit)."

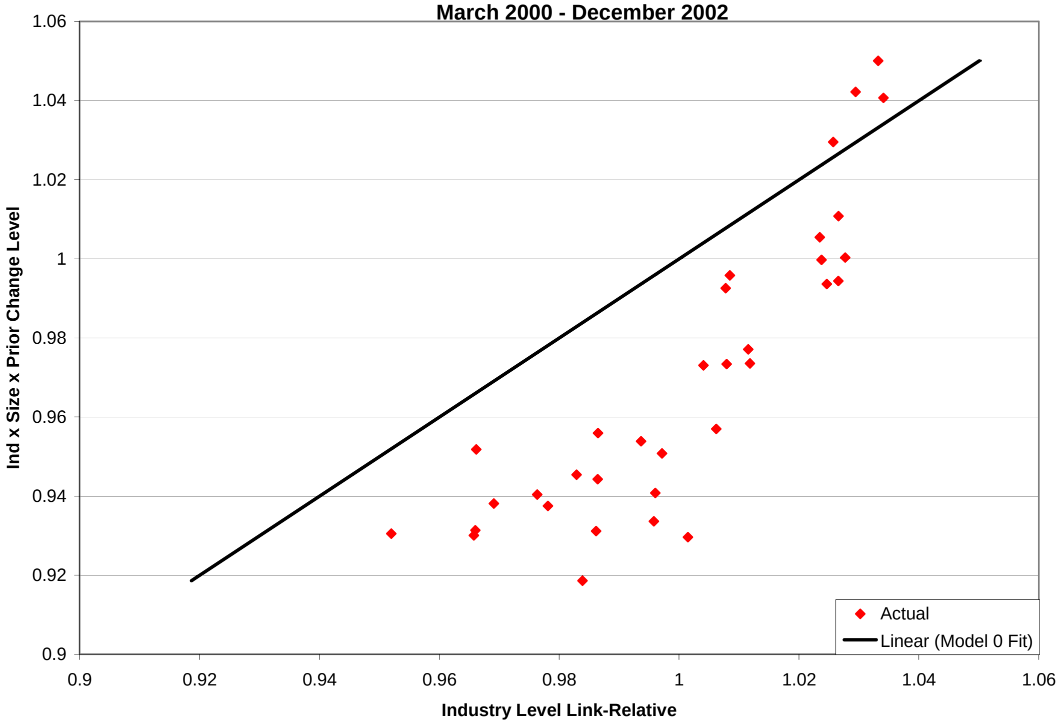
Figure 31-Comparison of Industry and Industry x Size x Change Link Relatives



Link-Relatives: Industry vs. Ind x Size x Prior Change
Construction: Prior Month Emp <10, Mid Prior Month Change
March 2000 - December 2002



Link-Relatives: Industry vs. Ind x Size x Prior Change
Construction: Prior Month Emp <10, High Prior Month Change
March 2000 - December 2002



For establishments in the Low and Mid prior month employment change classes, the link relatives for the industry as a whole tend to be below the actual link relatives (consistently so for the Low prior month employment change class), while the reverse is true for establishments in the High prior month employment change class. These results suggest the use of prior month employment size crossed with prior month employment change, at least for one or several smaller size classes, in Model 1 could better explain employment growth rates for potential late reporters than Model 0.

Based upon the preceding information, rough estimates of the bias in the current CES weighted link relative estimator under Model 1 were estimated, using the expected value derived in section A, and are presented in Table 35. Average bias refers to the bias derived using average values of δ_{tcg} and $\hat{p}^{(2)}_{(t-1)cg}$. Minimum bias refers to the bias estimated using average values of δ_{tcg} with minimum values of $\hat{p}^{(2)}_{(t-1)cg}$ if δ_{tcg} is positive and with maximum values of $\hat{p}^{(2)}_{(t-1)cg}$ if δ_{tcg} is negative. Maximum bias refers to the bias estimated using average values of δ_{tcg} with maximum values of $\hat{p}^{(2)}_{(t-1)cg}$ if δ_{tcg} is positive and with minimum values of $\hat{p}^{(2)}_{(t-1)cg}$ if δ_{tcg} is negative.

Table 35-Estimated Bias for Current Weighted Link Relative, $LR_{t,(t-1)c}$

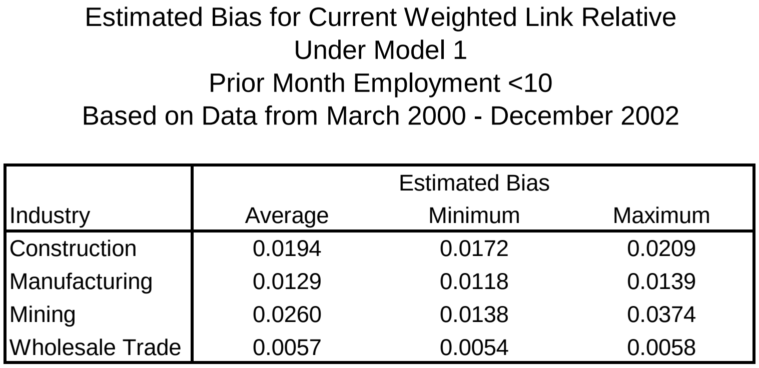
Estimated Bias for Current Weighted Link Relative
Under Model 1
Based on Data from March 2000 - December 2002

	Estimated Bias		
Industry	Average	Minimum	Maximum
Construction	0.0001	-0.0037	0.0032
Manufacturing	0.0000	-0.0009	0.0008
Mining	-0.0019	-0.0183	0.0061
Wholesale Trade	0.0001	-0.0020	0.0019

Results indicate small estimated biases on average for a monthly link relative although, given the estimate for a given month is linked to the benchmark through 11 to 23 months and the bias is cumulative, the estimated bias for a monthly employment estimate could be on the order of several tenths of a percentage point (and more than one percentage point for Mining. Given values for minimum and maximum estimated bias are fairly evenly balanced around zero, however, the biases could have a tendency to net out over time. In any given month there appears to be the potential for biases on the order of a tenth of a percentage point or more on the estimated link relative should the sample composition be skewed toward establishments with characteristics with lower growth rates than for the industry as a whole.

The estimated bias for small establishments, however, appears much more pronounced. Using the same approach, the estimated bias was derived for establishments with prior month employment <10. The results, provided in Table 36, show that estimated bias in the link relative for such establishments could be more than one percentage point. In addition, minimum and maximum estimated biases are not balanced around zero, and thus would tend to cumulate across time.

Table 36- Estimated Bias for Current Weighted Link Relative, $LR_{t,(t-1)c}$, when $Y_{(t-1)i} < 10$



Estimated Bias for Current Weighted Link Relative
Under Model 1
Prior Month Employment <10
Based on Data from March 2000 - December 2002

	Estimated Bias		
Industry	Average	Minimum	Maximum
Construction	0.0194	0.0172	0.0209
Manufacturing	0.0129	0.0118	0.0139
Mining	0.0260	0.0138	0.0374
Wholesale Trade	0.0057	0.0054	0.0058

In a similar fashion, the expected difference between preliminary and final estimates were derived, using values of δ_{tcg}, $\hat{p}^{(2)}_{(t-1)cg}$, and $\hat{p}^{(0)}_{(t-1)cg}$, and are presented in Table 37. Average bias refers to the bias derived using average values of δ_{tcg} and $\hat{p}^{(2)}_{(t-1)cg} - \hat{p}^{(0)}_{(t-1)cg}$. Minimum bias refers to the bias estimated using average values of δ_{tcg} with minimum values of $\hat{p}^{(2)}_{(t-1)cg} - \hat{p}^{(0)}_{(t-1)cg}$ if δ_{tcg} is positive and with maximum values of $\hat{p}^{(2)}_{(t-1)cg} - \hat{p}^{(0)}_{(t-1)cg}$ if δ_{tcg} is negative. Maximum bias refers to the bias estimated using average values of δ_{tcg} with maximum values of $\hat{p}^{(2)}_{(t-1)cg} - \hat{p}^{(0)}_{(t-1)cg}$ if δ_{tcg} is positive and with minimum values of $\hat{p}^{(2)}_{(t-1)cg} - \hat{p}^{(0)}_{(t-1)cg}$ if δ_{tcg} is negative.

Table 37-Estimated Revision for Preliminary Weighted Link Relative, $LR_{t,(t-1)c}$

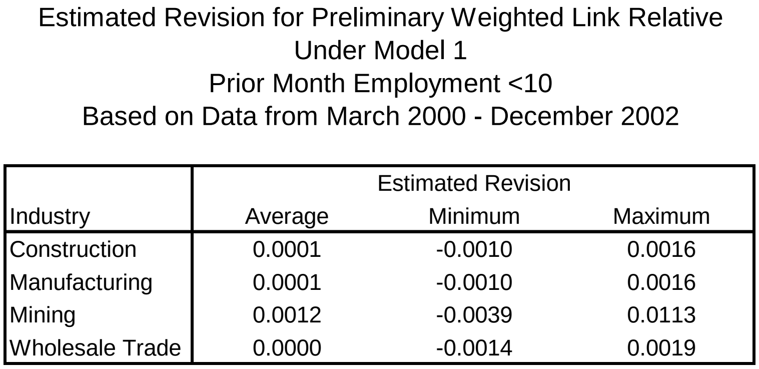
Estimated Revision for Preliminary Weighted Link Relative
Under Model 1
Based on Data from March 2000 - December 2002

	Estimated Revision		
Industry	Average	Minimum	Maximum
Construction	-0.0002	-0.0017	0.0009
Manufacturing	-0.0001	-0.0006	0.0003
Mining	0.0001	-0.0071	0.0072
Wholesale Trade	-0.0001	-0.0014	0.0015

Results indicate small estimated revisions on average for a monthly link. In any given month there appears to be the potential for revisions to the estimated link relative on the order of a tenth of a percentage point or more in either direction should the sample composition for preliminary reporters be skewed toward establishments with characteristics with lower growth rates than for the industry as a whole.

Estimated revisions for establishments with prior month employment <10 are also small, as indicated in Table 38. This is due to relatively small changes in the values for $\hat{p}_{(t-1)cg}$ between preliminary and final estimation, thus diminishing changes in the link relatives. Despite the potential bias for this subgroup discussed earlier, these results indicate the current weighted link relative estimator does not afford a reduction in that bias with increased sample reporting.

Table 38-Estimated Revision for Preliminary Weighted Link Relative, $LR_{t,(t-1)c}$, when $Y_{(t-1)i} < 10$



Estimated Revision for Preliminary Weighted Link Relative
Under Model 1
Prior Month Employment <10
Based on Data from March 2000 - December 2002

	Estimated Revision		
Industry	Average	Minimum	Maximum
Construction	0.0001	-0.0010	0.0016
Manufacturing	0.0001	-0.0010	0.0016
Mining	0.0012	-0.0039	0.0113
Wholesale Trade	0.0000	-0.0014	0.0019

C. Approach for Utilizing Incomplete Data

The results in the prior section suggest employment growth rate within industry is related to prior month employment size and prior month employment change, at least for establishments with small prior month employment. As the primary objective is to reduce differences between preliminary and revised estimates, the approach seeks to directly utilize information for sample establishments that subsequently become late reporters for month t. This can be accomplished through imputation of missing month t values for sample reporting in month $t-1$. While this approach results in the inclusion of sample units that do not subsequently become late reporters (i.e., that become nonresponders for month t), given late reporters make up the majority (~75%) of these sample units, it was felt this approach may yield smaller differences between preliminary and final estimates.

Imputation was utilized in the alternate approach rather than redefining estimation cells and carrying out a weighted link relative estimation at the refined cell level. Revising the definitions of estimation cells to incorporate the additional factors is not feasible, as population values for prior month employment size and change are not available on an ongoing basis. While a weighted link relative could be calculated at the refined cell level, the prior month estimated employment for the cell would be dynamic as establishments can change cells from month to month (i.e., the issue is values of $\hat{Y}_{(t-1)cg}$).

The approach developed here is intended to be used to impute for missing employment data due to sample units with reporting patterns resulting in missing data for month t when data are reported for month $t-1$

$$\begin{pmatrix}\mathbf{X}_{tci} & \mathbf{X}_{(t-1)ci}\end{pmatrix} = \begin{pmatrix} 0 & 1 \\ . & 0 \\ . & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ . & 1 \\ . & 0 \end{pmatrix}$$

and thus utilize, for preliminary estimates of month t employment, all sample units for which data were reported for month $t-1$.

1. Model 1: Proportional Growth Rate within Size and Prior Growth Class

The underlying model used for imputation assumes proportionality factors vary across classifications of establishments within industry

Model 1: $Y_{tcgi} = \rho_{tcg} Y_{(t-1)cgi} + k_{tcgi}$

$$k_{tcgi} \overset{ind}{\sim} N\left(0, \frac{\sigma_k^2 Y_{(t-1)cgi}}{w_{cgi}}\right)$$

where g represents the classification of sample unit i in industry c for month t based upon

$e_{(t-1)}$ (prior month employment size class)

$\Delta_{(t-1),(t-2)}$ (prior month employment change class)

Based on the results in section B, two sets of size classes were used in the evaluation: 1) <10 and 10+ (recognizing the distinctions in deviations seen at the size class level) – Model 1A; and 2) <10, 10-19, 20-49, and 50+ (recognizing potential additional distinctions in deviations seen at the size by employment growth class level) – Model 1B.

Table 39 contains the levels for prior employment size and prior employment change classes for Models 1A and 1B. For the large establishment size class, no further disaggregation by prior month employment change was made, given the results discussed previously.

Under this model the maximum likelihood estimator (MLE) for ρ_{tc} is

$$\hat{\rho}_{tcg} = \frac{\sum\limits_{i \in s_{cg}} w_{cgi} Y_{tcgi}}{\sum\limits_{i \in s_{cg}} w_{cgi} Y_{(t-1)cgi}}$$

As is done for the current CES weighted link relative estimator, estimates for ρ_{tcg} are derived using the set of constant reporters for months t and $t-1$. Analogous to the situation for model 0, these estimates will be model unbiased for $\hat{\rho}_{tcg}$ under Model 1.

Table 39-Cell Classifications within Industry for Model 1

Cell Classifications within Industry
Model 1A

Prior Month Employment	Designation for $e_{(t-1)}$	Prior Month Employment Change	Designation for $\Delta_{(t-1),(t-2)}$
<10	1	Low Third	L
		Mid Third	M
		Top Third	H
		Unknown	U
10+	4	n/a	-

Model 1B

Prior Month Employment	Designation for $e_{(t-1)}$	Prior Month Employment Change	Designation for $\Delta_{(t-1),(t-2)}$
<10	1	Low Third	L
		Mid Third	M
		Top Third	H
		Unknown	U
10-19	2	Low Third	L
		Mid Third	M
		Top Third	H
		Unknown	U
20-49	3	Low Third	L
		Mid Third	M
		Top Third	H
		Unknown	U
50+	4	n/a	-

Model 1 assumes values of ρ_{tcg} are the same for month t reporters and nonreporters (i.e., expected growth rate is the same for late reporters and preliminary reporters within a cell/class cg). A first question is the appropriateness of this assumption.

If Model 1 provides a good description of the population distribution, then the difference between link relatives for preliminary and late reporters should be small. Table 40 contains comparisons of deviations in link relatives between preliminary and late reporters for redefined cells versus industry level. These results show greater comparability of link relatives associated with the redefined cell definitions, as both the average deviation and the average absolute deviation for the redefined cells are generally lower than the corresponding deviations for the industry level.

Table 40-Deviations of Link Relatives, $LR_{t,(t-1)c}$: Preliminary vs. Late Reporters

Diagnostics for Fit of Link-Relatives
Industry vs. Industry x Prior Month Employment x Prior Month Employment Change
March 2000 - December 2002

Industry	Prior Month Employment Size	Prior Month Employment Change	Average Sample Size		Deviation of Link-relatives - Preliminary vs. Late Reporter			
					Ind x Prior Month Emp x Prior Month Emp Change Level		Industry Level	
			Preliminary Reporters	Late Reporters	Average	Ave Abs	Average	Ave Abs
Construction	<10	Low	404	66	0.0425	0.0804	0.1221	0.1393
		Mid	1521	210	0.0025	0.0195	-0.0144	0.0381
		High	322	51	-0.0376	0.0464	-0.0903	0.0931
		Unk	97	21	-0.0061	0.0859	-0.0035	0.0870
	10-19	Low	191	28	0.0128	0.0781	0.0339	0.0806
		Mid	669	90	-0.0042	0.0172	-0.0355	0.0386
		High	196	30	-0.0135	0.0480	-0.0654	0.0798
		Unk	30	6	0.0418	0.1186	0.0049	0.1161
	20-49	Low	308	46	0.0025	0.0540	-0.0187	0.0572
		Mid	922	135	0.0017	0.0152	-0.0303	0.0320
		High	306	43	0.0066	0.0336	-0.0182	0.0529
		Unk	34	9	-0.0060	0.0897	-0.0456	0.0752
	50+	n/a	2883	490	0.0014	0.0128	-0.0042	0.0122
Manufacturing	<10	Low	200	56	0.0574	0.0800	0.1423	0.1439
		Mid	901	173	0.0034	0.0258	0.0045	0.0254
		High	126	33	0.0002	0.0510	-0.0370	0.0537
		Unk	50	15	0.0453	0.1351	0.0765	0.1306
	10-19	Low	234	44	0.0080	0.0373	0.0326	0.0419
		Mid	556	93	-0.0073	0.0148	-0.0110	0.0155
		High	223	44	-0.0040	0.0244	-0.0174	0.0301
		Unk	29	9	0.0192	0.0715	0.0214	0.0626
	20-49	Low	380	64	0.0055	0.0201	0.0216	0.0302
		Mid	1204	203	-0.0039	0.0084	-0.0080	0.0104
		High	349	58	-0.0003	0.0204	-0.0067	0.0230
		Unk	50	13	-0.0166	0.0579	-0.0125	0.0540
	50+	n/a	11537	2512	-0.0015	0.0050	-0.0018	0.0043
Mining	<10	Low	36	9	-0.0052	0.3140	0.1742	0.3366
		Mid	167	35	0.0248	0.0827	0.0309	0.0789
		High	24	6	0.0644	0.2954	0.0624	0.2766
		Unk	7	3	-0.3468	0.4392	-0.2652	0.3476
	10-19	Low	37	7	0.0619	0.2524	0.0979	0.2412
		Mid	102	24	0.0107	0.0588	-0.0026	0.0521
		High	33	6	-0.0505	0.1561	-0.0734	0.1497
		Unk	4	3	-0.4836	0.6184	-0.5161	0.5293
	20-49	Low	42	11	-0.0157	0.0855	0.0017	0.0736
		Mid	151	36	0.0066	0.0305	0.0087	0.0319
		High	54	15	0.0095	0.0850	-0.0055	0.0768
		Unk	6	3	-0.2227	0.4099	-0.2324	0.2849
	50+	n/a	361	115	0.0004	0.0131	0.0004	0.0100
Wholesale Trade	<10	Low	164	48	0.0131	0.0716	0.0635	0.0927
		Mid	1419	387	0.0005	0.0133	-0.0018	0.0137
		High	116	31	-0.0157	0.0648	-0.0423	0.0722
		Unk	54	33	0.0339	0.0819	0.0411	0.0671
	10-19	Low	152	42	-0.0096	0.0445	0.0045	0.0388
		Mid	478	116	-0.0052	0.0149	-0.0102	0.0159
		High	151	37	0.0018	0.0253	-0.0174	0.0304
		Unk	20	14	-0.0632	0.1127	-0.0310	0.0607
	20-49	Low	118	39	0.0012	0.0362	0.0140	0.0258
		Mid	537	160	0.0011	0.0082	-0.0053	0.0087
		High	171	48	-0.0009	0.0194	-0.0157	0.0237
		Unk	24	15	-0.0603	0.0968	-0.0693	0.0936
	50+	n/a	1483	537	-0.0015	0.0045	-0.0024	0.0042

A second question is specification of the assumed distribution for units other than those reporting in month $t-1$ (i.e., other than constant reporters or units for which imputations are carried out). As stated previously, the common approach of estimating link relatives for each of the classes and multiplying by the prior month's estimate is not valid as the population within a class changes over time and thus, an estimate of the population value is not available.

An alternative is to take a pattern-mixture model (Little, 1993) approach. The population can be assumed to be divided into three groups:

1) Units for which data for both months t and $t-1$ are available. These are the units currently used in the weighted link relative.

2) Units for which data for only month $t-1$ are available. These are the units for which the alternative approach will derive imputed values for use in the weighted link relative.

3) Units for which data for month $t-1$ are not available. These represent a combination of nonsampled, nonreporters for both months t and $t-1$, and units reporting for month t but not month $t-1$.

Each of these three groups of units has a different missing data pattern. Under the pattern-mixture model approach, growth rate is assumed dependent upon missing data pattern, e.g.,

$$Y_{tMcgi} = \rho_{tMcg} Y_{(t-1)Mcgi} + k_{tMcgi}$$

$$k_{tcMgi} \overset{ind}{\sim} N\left(0, \frac{\sigma_k^2 Y_{(t-1)cMgi}}{w_{cMgi}}\right)$$

where M refers to missing data pattern as defined above.

Growth rates for missing data patterns 2 and 3 cannot be estimated from the data. Therefore identifying restrictions linking the parameters for the models for missing data patterns 2 and 3 are linked to those for missing data pattern 1 so as to allow estimation of parameters. The identifying restrictions for missing data pattern 2 assumes equivalence of growth rates within the redefined cells

$$\rho_{t2cg} = \rho_{t1cg}$$

This identifying restriction allows imputation of missing values based upon the estimated growth rates for a cell based upon constant reporters.

For missing data pattern 3, the intention is to use the weighted link relative within an industry based upon the set of constant reporters plus reporters for month $t-1$` with imputed values to estimate the link relative for the industry. This assumes the identifying restriction for missing data pattern 3 links the growth for units in the missing data pattern at the industry level to the marginal (at the industry level) of the growth rates for missing data patterns 1 and 2

$$\rho_{t3c} = \rho_{t.c.}$$

This marginal can be derived by taking the expected value of the weighted link relative utilizing data from missing data patterns 1 and 2 under Model 1.

$$\rho_{t.c.} = E\left[\frac{\sum_{g}\sum_{\mathrm{M}=1,2}\sum_{\mathrm{i}\in \mathrm{M}_1,\mathrm{M}_2} \mathrm{w}_{\mathrm{tMcgi}}\mathrm{Y}_{\mathrm{tMcgi}}}{\sum_{g}\sum_{\mathrm{M}=1,2}\sum_{\mathrm{i}\in \mathrm{M}_1,\mathrm{M}_2} \mathrm{w}_{\mathrm{tMcgi}}\mathrm{Y}_{(\mathrm{t}\text{-}1)\mathrm{Mcgi}}} \mid \mathrm{Model\,1}\right] = \left[\frac{\sum_{g}\sum_{\mathrm{M}=1,2}\sum_{\mathrm{i}\in \mathrm{M}_1,\mathrm{M}_2} \mathrm{w}_{\mathrm{tMcgi}}\mathrm{E}\left(\mathrm{Y}_{\mathrm{tMcgi}} \mid \mathrm{Model\,1}\right)}{\sum_{g}\sum_{\mathrm{M}=1,2}\sum_{\mathrm{i}\in \mathrm{M}_1,\mathrm{M}_2} \mathrm{w}_{\mathrm{tMcgi}}\mathrm{Y}_{(\mathrm{t}\text{-}1)\mathrm{Mcgi}}}\right]$$

$$= \left[\frac{\sum_{g}\sum_{\mathrm{M}=1,2}\sum_{\mathrm{i}\in \mathrm{M}_1,\mathrm{M}_2} \mathrm{w}_{\mathrm{tMcgi}}\rho_{\mathrm{tMcg}}\mathrm{Y}_{(\mathrm{t}\text{-}1)\mathrm{Mcgi}}}{\sum_{g}\sum_{\mathrm{M}=1,2}\sum_{\mathrm{i}\in \mathrm{M}_1,\mathrm{M}_2} \mathrm{w}_{\mathrm{tMcgi}}\mathrm{Y}_{(\mathrm{t}\text{-}1)\mathrm{Mcgi}}}\right] = \left[\frac{\sum_{g}\rho_{\mathrm{t1cg}}\sum_{\mathrm{M}=1,2}\sum_{\mathrm{i}\in \mathrm{M}_1,\mathrm{M}_2} \mathrm{w}_{\mathrm{tMcgi}}\mathrm{Y}_{(\mathrm{t}\text{-}1)\mathrm{Mcgi}}}{\sum_{g}\sum_{\mathrm{M}=1,2}\sum_{\mathrm{i}\in \mathrm{M}_1,\mathrm{M}_2} \mathrm{w}_{\mathrm{tMcgi}}\mathrm{Y}_{(\mathrm{t}\text{-}1)\mathrm{Mcgi}}}\right]$$

$$= \sum_{g} \rho_{\mathrm{t1cg}} \hat{p}_{tcg}$$

where $\hat{p}_{tcg}$ is an estimate of the proportion of the population total for estimation cell c within class g.

Note that this is similar to the expected value of the current weighted link relative under Model 1 since $\rho_{tcg} = \rho_{tc} + \delta_{tcg}$. The difference from the previous result is that $\hat{p}_{tcg}$ is based on all month $t-1$ reporters instead of just constant reporters for both months t and $t-1$.

2. Model 2: Stable Effect of Prior Month Employment Change within Size Class

Information on δ_{tcg} presented in section B suggests the effect of prior employment growth rate for establishments with prior month employment <10 may be relatively stable. As a result, an alternative to Model 1, assuming the effect of prior employment change does not depend on time t, was also considered. This model used the size classes from Model 1A. This model can be written as

$$\text{Model 2: } Y_{tce_{(t-1)}\Delta i} = \begin{cases} \left(\rho_{tce_{(t-1)}} + \boldsymbol{\lambda}^{T}_{ce_{(t-1)}} \boldsymbol{\Delta}_{(t-1),(t-2)ce_{(t-1)}i}\right) Y_{(t-1)ce_{(t-1)}\Delta i} + k_{tce_{(t-1)}\Delta i}, \; e_{(t-1)} < 10 \\ \rho_{tce_{(t-1)}} Y_{(t-1)ce_{(t-1)}i} + k_{tce_{(t-1)}i}, \; e_{(t-1)} = 10+ \end{cases}$$

$$k_{tce_{(t-1)}\Delta i} \overset{ind}{\sim} N\left(0, \frac{\sigma_k^2 Y_{(t-1)ce_{(t-1)}\Delta i}}{w_{ce_{(t-1)}\Delta i}}\right)$$

where $\rho_{tce_{(t-1)}}$ is the underlying employment growth rate from month $t-1$ to month t

for industry c and prior employment size class $e_{(t-1)}$

$\mathbf{\Delta}_{(t-1),(t-2),ce_{(t-1)}i} = \begin{pmatrix} \Delta^{(low)}_{(t-1),(t-2),ce_{(t-1)}i} \\ \Delta^{(mid)}_{(t-1),(t-2),ce_{(t-1)}i} \\ \Delta^{(high)}_{(t-1),(t-2),ce_{(t-1)}i} \\ \Delta^{(unk)}_{(t-1),(t-2),ce_{(t-1)}i} \end{pmatrix}$ is the vector containing all 0's and a single 1 to designate prior month's employment change for sample units in the small size class, rank ordered into four groups – low, middle, high, unknown (i.e., employment change not reported for months $t-1$ and/or $t-2$)

Note that the $\rho_{tce_{(t-1)}}$ and $\boldsymbol{\lambda}_{ce_{(t-1)}}$ are fixed effects at any given time period; however, $\boldsymbol{\lambda}$ does not depend upon month and therefore can be estimated using data from previous months. Estimation of the $\boldsymbol{\lambda}_{ce_{(t-1)}}$ for Model 2 was carried out using Bayes' estimation with data for the six months prior to month t, as described in section D. Estimation of the $\rho_{ce_{(t-1)}}$ was carried out using weighted link relatives for the constant reporters for months t and $t-1$, in the same manner as the growth rates for Model 1.

In practice the dimension of the $\boldsymbol{\lambda}$ and $\mathbf{\Delta}$ were reduced by one, since any one element is linearly dependent on the remaining elements. The element selected for exclusion from the vector becomes the reference level for the factor. The mid group $\left(\Delta^{(mid)}\right)$ was selected to be the reference level.

D. Empirical Analysis of Model Performance

Estimates generated using a completed dataset consisting of reported data and data imputed using Models 1A, 1B, and 2 were compared to those generated using reported data only (current method – Model 0) for the period March 2000 through

December 2002 (April 2001 through March 2002 for Model 2). Preparation of the CES data for the research analysis was described in Chapter III.

Statistics of interest were total employment for the month and the change in total employment from the prior month. Performance assessment was made on the basis of revisions between preliminary $(k=0)$ and final $(k=1)$ estimates. In addition, estimates for March 2001 and 2002 (referred to below as the "benchmark" months, t_B) were compared with the total employment from the ES-202 program.

1. Generating Estimates

Employment estimates were generated using the current CES link relative estimator. A separate dataset was created for each approach – a dataset consisting of reported data only and three datasets consisting of reported data plus data imputed for late reporters and nonrespondents using Models 1A, 1B, and 2.

For each data set, two sets of estimates were generated for each month – preliminary and final – using SAS v.8.2. The fixed effects for Model 2 were estimated using a Bayes' approach, described in Appendix G. For the dataset consisting of reported data only, preliminary estimates were based upon those reporting by the preliminary cutoff date, d_t, for month t, while final estimates were based on those reporting by the final cutoff date. For the completed datasets, preliminary estimates were based upon data reported by d_t, plus data imputed for late reporters and nonrespondents, while final estimates were based upon data reported by the final cutoff date plus data imputed for nonrespondents (i.e., imputed data was replaced with reported data for late reporters, while imputed data remained the same for nonrespondents). The SAS code used in deriving the imputed values and calculating the link relatives is provided in Appendix H.3.

2. Variance Estimates

Variance estimates for weighted link relatives were derived using the CES BRR method described in Chapter III. As discussed in Shao, et al. (1998), imputing for missing values separately for each replicate based on data within the appropriate half-sample recovers variance due to imputation and produces consistent variance estimates for a class of estimators that are smooth functions of totals, which encompasses the weighted link relative.

This approach to variance estimation was carried out for the empirical analysis. Model coefficients were estimated separately for each replicate and half-sample. The one exception was that the fixed effects coefficients for Model 2 were not reestimated for each replicate, due to length of time required for computing. As a result, the errors presented will underestimate the total errors associated with link relatives from Model 2. The SAS code used in calculating the half-sample estimates is provided in Appendix H.4.

3. Measures of Accuracy

This dissertation research was carried out to develop an estimator for employment in the CES survey that would result in a reduction in the magnitude of revisions between preliminary and final estimates of monthly employment and month-to-month change in employment. Assessment of the performance of the proposed estimator can be made by comparison to final estimates.

Monthly estimates of the link relatives and associated standard deviations for the approaches are provided in Appendix I. One item of note is the size of the standard deviations associated with the estimated link relatives. As seen in Table 41, the standard deviations dominate revisions between preliminary and final estimates. This will limit the conclusions that can be drawn from the analysis to observations.

Table 41-Summary Information for Estimated Link Relatives, $LR_{t,(t-1)c}$

Average Revisions, Standard Deviations for Estimated Link Relatives
March 2000 - December 2002

	Current			Model 1A			Model 1B			Model 2		
	Average st dev		Average Absolute Revision	Average st dev		Average Absolute Revision	Average st dev		Average Absolute Revision	Average st dev		Average Absolute Revision
Industry	Preliminary	Final		Preliminary	Final		Preliminary	Final		Preliminary	Final	
Construction	0.0106	0.0107	0.0011	0.0106	0.0107	0.0010	0.0108	0.0108	0.0010	0.0115	0.0118	0.0010
Manufacturing	0.0033	0.0033	0.0009	0.0033	0.0033	0.0008	0.0033	0.0033	0.0008	0.0034	0.0036	0.0010
Mining	0.0142	0.0132	0.0029	0.0143	0.0130	0.0029	0.0141	0.0129	0.0029	0.0120	0.0097	0.0025
Wholesale Trade	0.0064	0.0055	0.0008	0.0065	0.0056	0.0008	0.0066	0.0056	0.0008	0.0065	0.0060	0.0009

Monthly estimates of employment were derived by utilizing March 2002 ES-202 data as the benchmark month, and moving the estimates forward by multiplying link relatives across months. Preliminary estimates were calculated as the preliminary link relative times the prior month's final estimate of employment.

For monthly estimates, the performance measure used is the relative revision between preliminary and final estimates

$$\text{Rel Rev}_{0|1}\left(\hat{Y}_t^{(0)}\right)=\frac{\hat{Y}_t^{(2)}-\hat{Y}_t^{(0)}}{\hat{Y}_t^{(0)}}$$

The difference in absolute relative revisions between that for the current method and that for an alternative method provides an indication of the reduction in the magnitude of the revision. Table 42 provides summary information for the relative revisions across the period April 2000 through December 2002. Revisions for alternative methods are essentially the same as those for the current method, although the alternative methods achieved a slight reduction in the average revision.

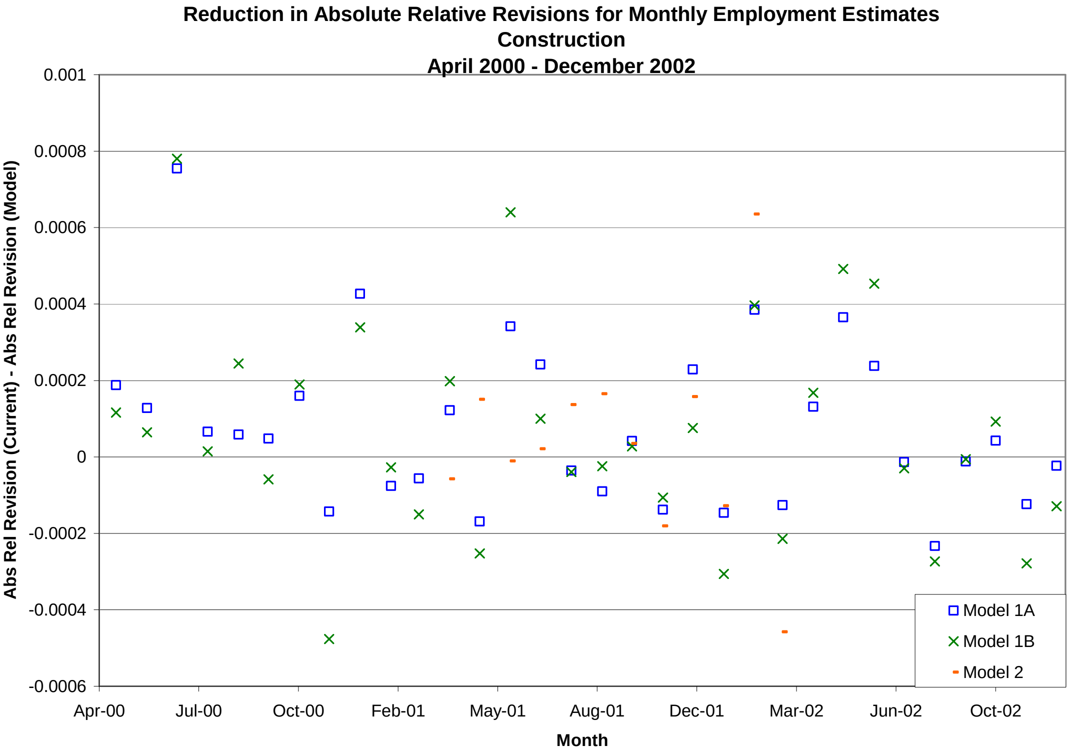
Table 42-Relative Revisions for Monthly Employment Estimates, $\hat{Y}_t$

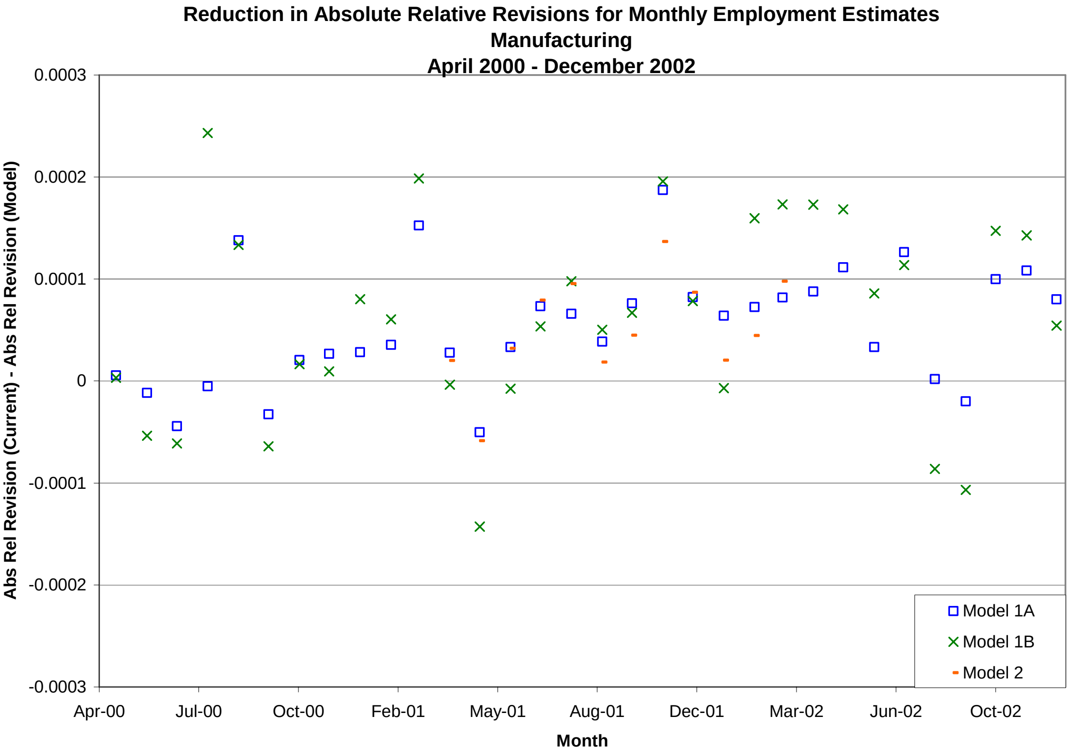
Relative Revisions in Estimated Monthly Employment
April 2000 - December 2002

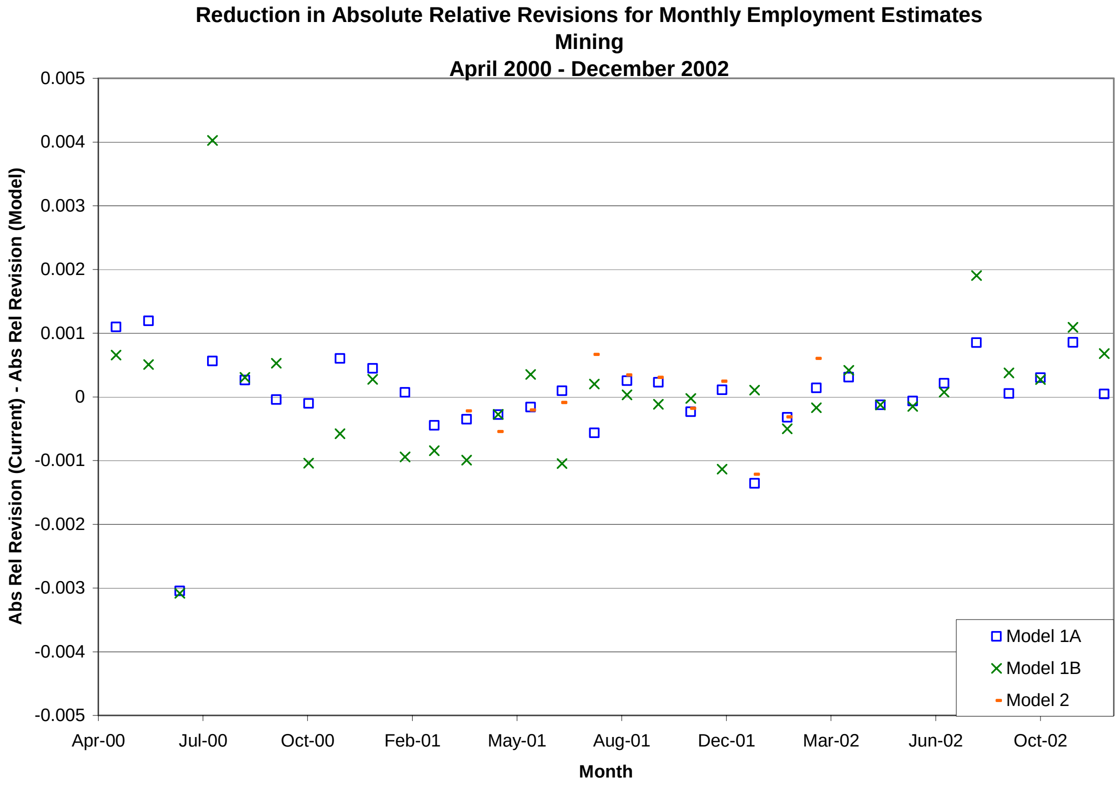
Industry	Metric	Current	Model 1A	Model 1B	Model 2
Construction	Average Revision	0.00%	0.01%	0.02%	0.02%
	Average Absolute Revision	0.11%	0.10%	0.10%	0.10%
	Average Reduction in Absolute Revision	-	0.01%	0.01%	0.00%
Manufacturing	Average Revision	-0.02%	-0.02%	-0.02%	-0.01%
	Average Absolute Revision	0.08%	0.08%	0.08%	0.10%
	Average Reduction in Absolute Revision	-	0.01%	0.01%	0.01%
Mining	Average Revision	0.04%	0.05%	0.04%	0.14%
	Average Absolute Revision	0.30%	0.30%	0.30%	0.25%
	Average Reduction in Absolute Revision	-	0.00%	0.00%	-0.01%
Wholesale Trade	Average Revision	-0.02%	-0.01%	0.00%	-0.03%
	Average Absolute Revision	0.08%	0.08%	0.08%	0.09%
	Average Reduction in Absolute Revision	-	0.01%	0.00%	0.02%

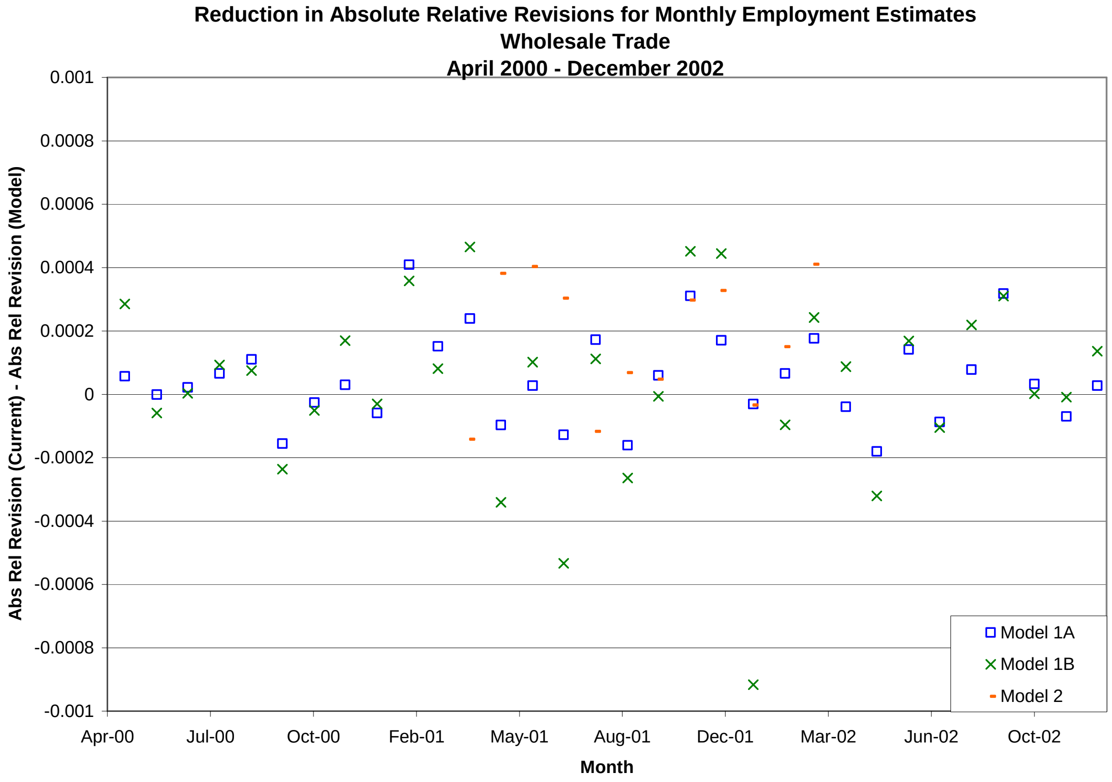
The distributions for the reductions in absolute relative revisions are plotted in Figure 32. A positive value in the figure means that the revisions are smaller under the model that with the current method. The graphs suggest a general tendency for the magnitude of the relative revisions for the alternate approaches to be less than the relative revisions for the current method in Manufacturing and Wholesale Trade.

Figure 32-Reduction in Absolute Relative Revision for Monthly Employment Estimates, $\hat{Y}_t$






Reduction in Absolute Relative Revisions for Monthly Employment Estimates
Mining
April 2000 - December 2002
Abs Rel Revision (Current) - Abs Rel Revision (Model)
0.005
0.004
0.003
0.002
0.001
0
-0.001
-0.002
-0.003
-0.004
-0.005
Apr-00
Jul-00
Oct-00
Feb-01
May-01
Aug-01
Dec-01
Mar-02
Jun-02
Oct-02
Month
□ Model 1A
× Model 1B
- Model 2

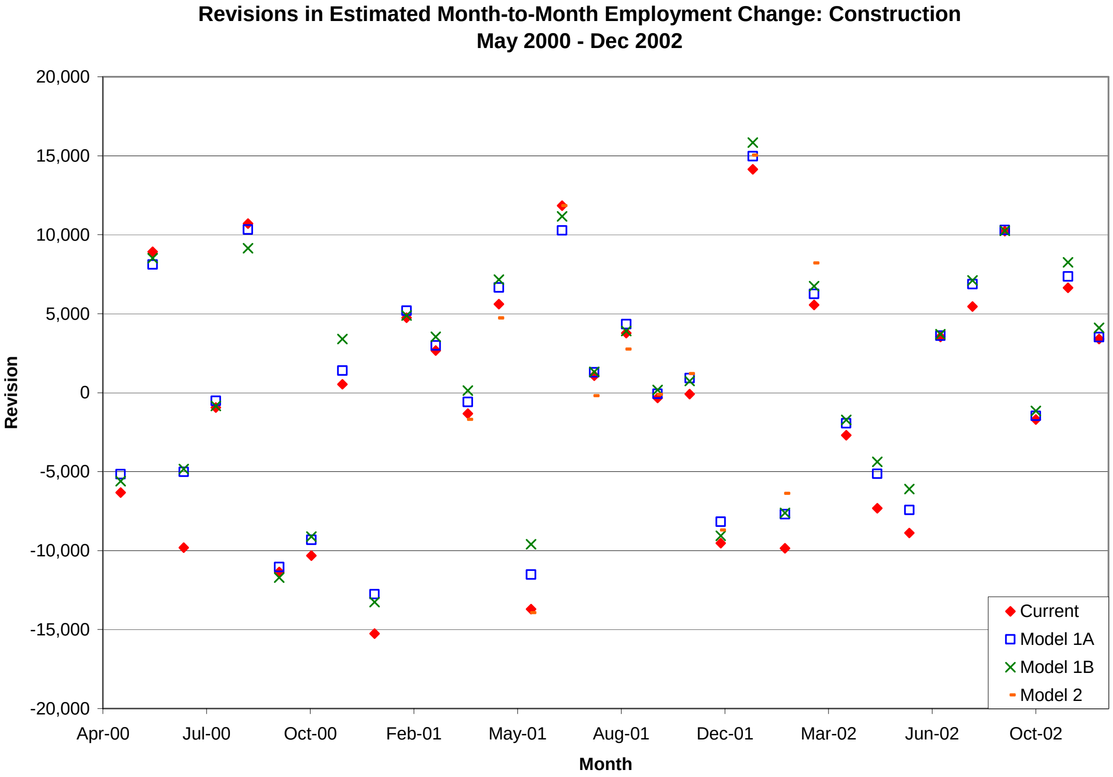

Reduction in Absolute Relative Revisions for Monthly Employment Estimates
Wholesale Trade
April 2000 - December 2002
Abs Rel Revision (Current) - Abs Rel Revision (Model)
0.001
0.0008
0.0006
0.0004
0.0002
0
-0.0002
-0.0004
-0.0006
-0.0008
-0.001
Apr-00
Jul-00
Oct-00
Feb-01
May-01
Aug-01
Dec-01
Mar-02
Jun-02
Oct-02
Month
□ Model 1A
× Model 1B
- Model 2

For estimates of month-to-month change, the performance measure used is the actual revision between preliminary and final estimates

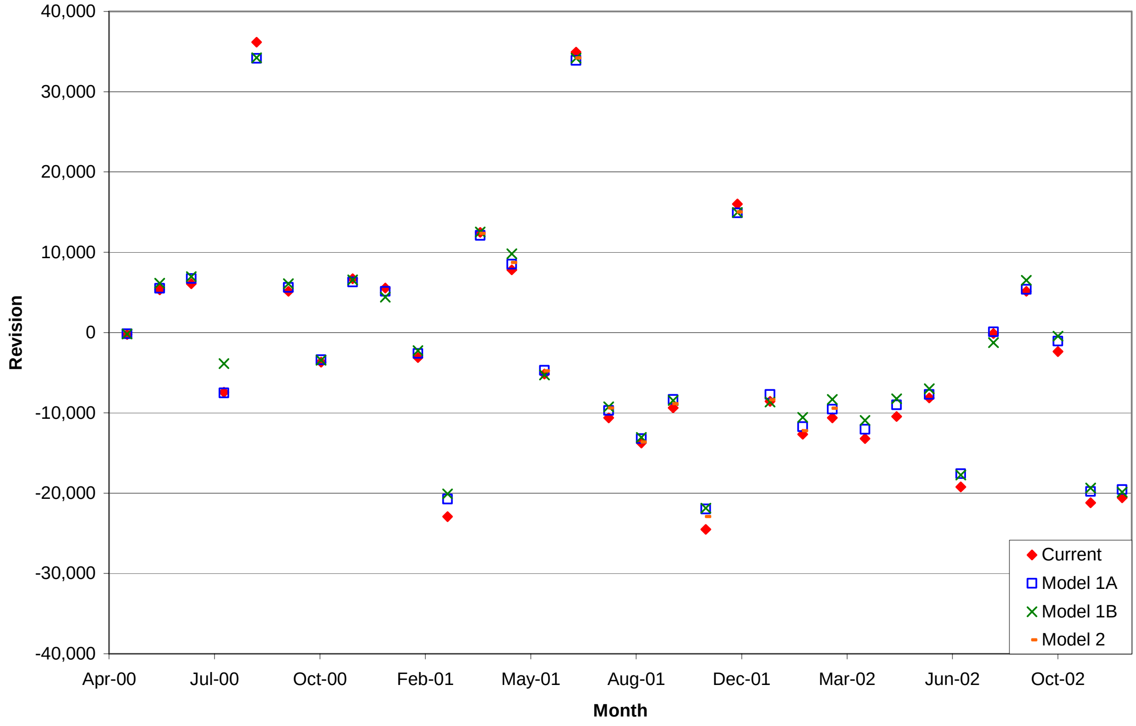
$$\text{Rev}_{0|1}\left(\hat{\Delta}^{(0)}_{t,t-1}\right)=\hat{\Delta}^{(2)}_{t,t-1}-\hat{\Delta}^{(0)}_{t,t-1}$$

Revisions in month-to-month change estimates are graphed in Figure 33. There appears to be a general tendency for larger revisions in month-to-month change for the current method versus the alternative methods, especially related to larger month-to-month change estimates.

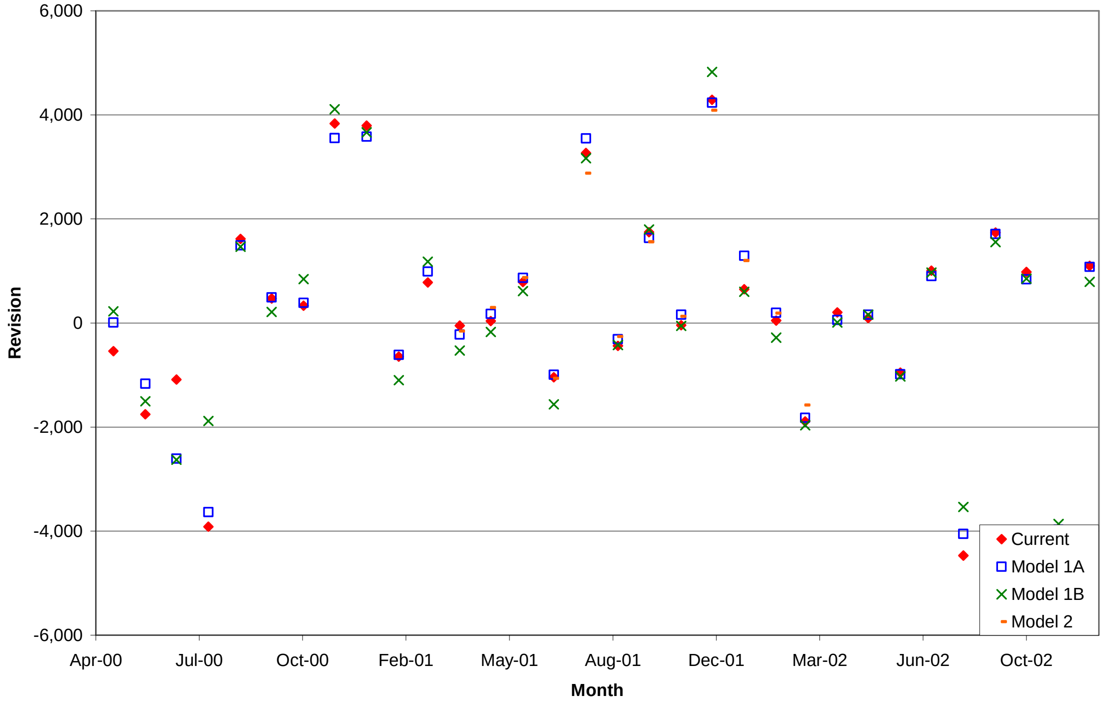
Figure 33-Revisions in Month-to-Month Change Estimates, $\hat{\Delta}_{t,(t-1)}$

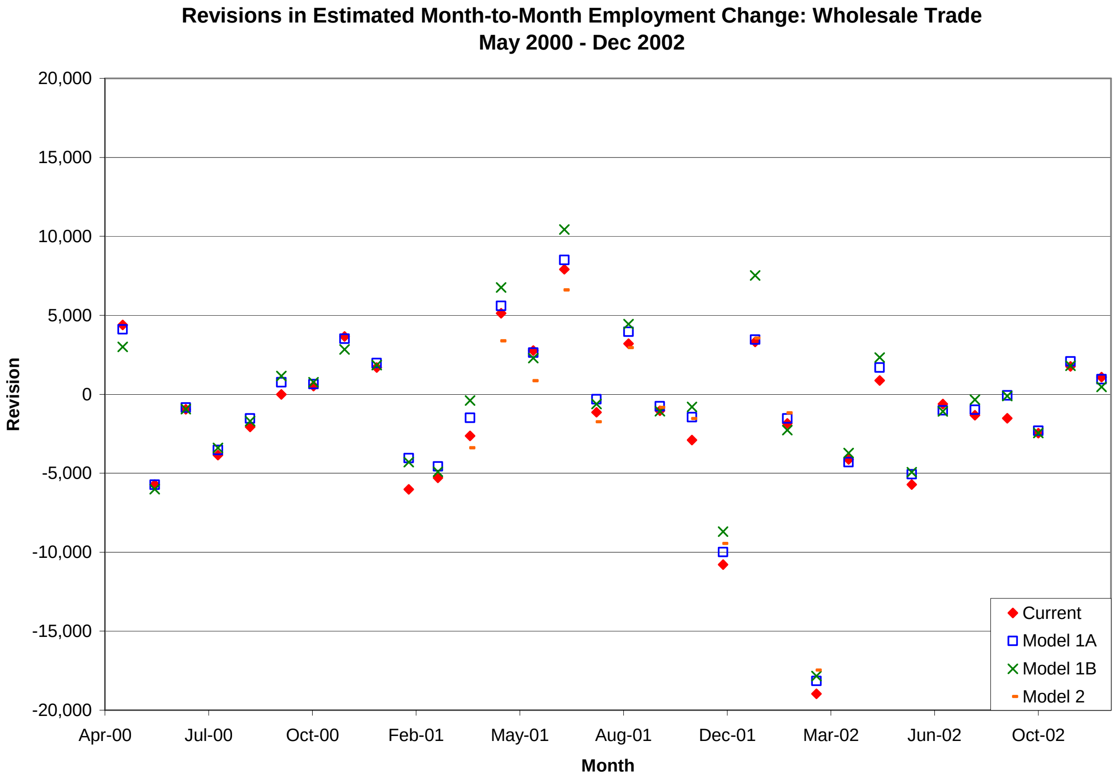


Revisions in Estimated Month-to-Month Employment Change: Manufacturing May 2000 - Dec 2002



Revisions in Estimated Month-to-Month Employment Change: Mining May 2000 - Dec 2002





For all industries, there is a reduction in the absolute revision of month-to-month change estimates, on average across the months, as seen in Table 43. This reduction, although less than 1,000 on average, does represent 6% - 8% of the average revision for the current method.

Table 43-Summary of Revisions in Month-to-Month Change Estimates, $\hat{\Delta}_{t,(t-1)}$

Absolute Revision in Estimated Month-to-Month Change in Employment
May 2000 - December 2002

Industry	Metric	Current	Model 1A	Model 1B	Model 2
Construction	Average Absolute Revision	6,506	6,006	6,095	6,232
	Average Reduction in Absolute Revision	-	500	411	273
Manufacturing	Average Absolute Revision	11,540	10,824	10,702	13,327
	Average Reduction in Absolute Revision	-	716	837	-1,787
Mining	Average Absolute Revision	1,500	1,492	1,487	1,190
	Average Reduction in Absolute Revision	-	7	13	310
Wholesale Trade	Average Absolute Revision	3,603	3,362	3,476	4,412
	Average Reduction in Absolute Revision	-	241	128	-809

At a more local level, the performance of the model can be evaluated by comparing imputed values to actual values for late reporters. Imputation error for a set of late reporters can be defined as

$$\text{Rel Err(Method m)} = \frac{\sum_{X_{ti}^{LR}=1} \hat{Y}_{ti,m} - \sum_{X_{ti}^{LR}=1} Y_{ti}}{\sum_{X_{ti}^{LR}=1} Y_{ti}}$$

where Y_{ti} represents the month t reported employment from sample establishment i

$\hat{Y}_{ti,m}$ represents the imputed employment for month t for sample establishment i, based on imputation method m

Note that for the current weighted link relative estimator, the imputed employment for a sample establishment is equal to the prior month employment for that establishment times the preliminary link relative for the corresponding estimation cell.

$$\hat{Y}_{tci,m} = LR^{(0)}_{t,(t-1)c} Y_{(t-1)ci}$$

Table 44 contains summary information on average relative errors by prior month size class, and by prior month employment change within prior month size class, for the period March 2000 – December 2002. Both 10+ and 10-19, 20-49, 50+ size classes are shown, with the results for Model 1 based upon the corresponding Model 1A (10+) or Model 1B (10-19, 20-49, 50+). These data show the reduction in errors for establishments with prior month employment <10, especially those with Low prior month employment change. These data also indicate that improvements due to use of Model 1 are fairly well restricted to establishments with prior month employment size <10.

Table 44-Relative Errors in Predicting Employment for Late Reporters

Relative Errors in Predicting Employment for Late Reporters
March 2000 - December 2002

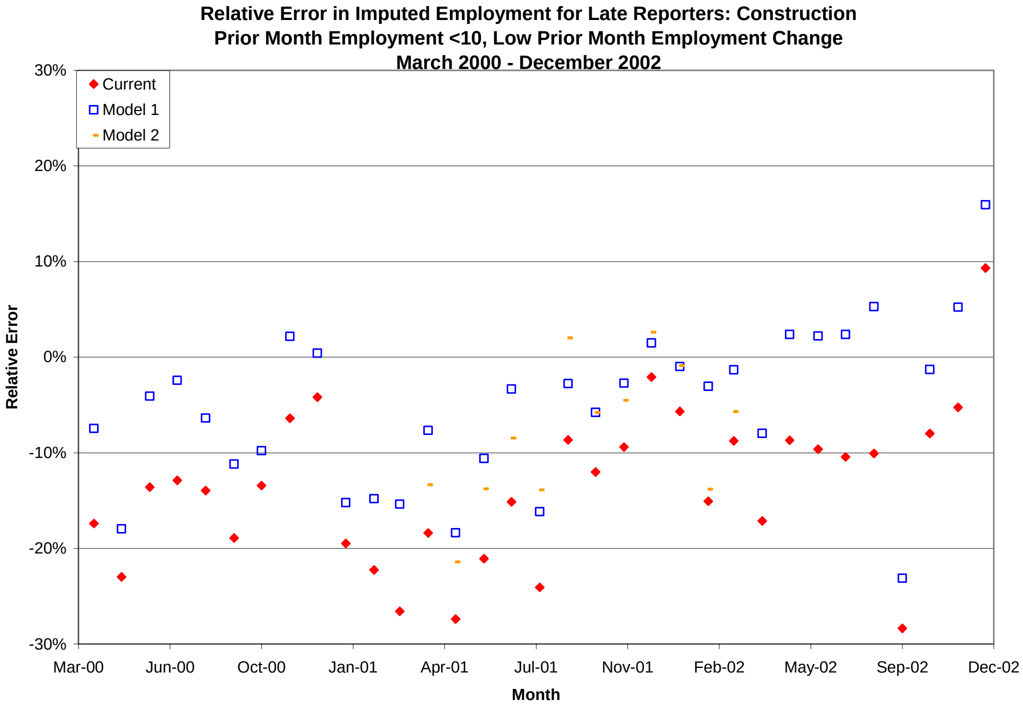
Size Class	Metric	Construction			Manufacturing			Mining			Wholesale Trade		
		Current	Model 1	Model 2	Current	Model 1	Model 2	Current	Model 1	Model 2	Current	Model 1	Model 2
<10	Average Relative Error	-5.5%	-3.3%	-3.8%	-5.5%	-3.9%	-4.8%	-7.8%	-5.9%	-5.5%	-1.0%	-0.3%	0.7%
	Average Absolute Relative Error	5.6%	3.9%	3.8%	5.7%	4.4%	5.1%	9.8%	8.7%	9.7%	1.4%	1.1%	2.0%
	Average Reduction in Absolute Relative Error		1.7%	2.2%		1.2%	1.8%		1.2%	0.5%		0.3%	-0.7%
10+	Average Relative Error	0.1%	-0.3%	-0.3%	0.0%	0.0%	-0.1%	-0.3%	-0.4%	-0.6%	0.1%	0.0%	-0.2%
	Average Absolute Relative Error	0.8%	0.8%	0.9%	0.3%	0.3%	0.3%	1.1%	1.2%	1.2%	0.4%	0.4%	0.5%
	Average Reduction in Absolute Relative Error		0.0%	-0.1%		0.0%	0.0%		0.0%	-0.1%		0.0%	0.0%
10-19	Average Relative Error	-1.7%	-1.7%	n/a	-0.3%	0.0%	n/a	-0.3%	-1.4%	n/a	0.3%	0.5%	n/a
	Average Absolute Relative Error	2.2%	2.5%	n/a	1.8%	1.8%	n/a	3.9%	4.4%	n/a	1.2%	1.3%	n/a
	Average Reduction in Absolute Relative Error		-0.2%	n/a		0.0%	n/a		-0.5%	n/a		-0.1%	n/a
20-49	Average Relative Error	-1.7%	-1.9%	n/a	-0.6%	-0.5%	n/a	0.3%	-0.2%	n/a	0.3%	0.2%	n/a
	Average Absolute Relative Error	2.0%	2.1%	n/a	1.1%	1.1%	n/a	2.0%	2.3%	n/a	0.9%	1.1%	n/a
	Average Reduction in Absolute Relative Error		-0.1%	n/a		0.0%	n/a		-0.3%	n/a		-0.2%	n/a
50+	Average Relative Error	0.3%	-0.3%	n/a	0.0%	0.0%	n/a	-0.3%	-0.2%	n/a	0.1%	-0.1%	n/a
	Average Absolute Relative Error	0.9%	0.9%	n/a	0.3%	0.3%	n/a	1.2%	1.4%	n/a	0.4%	0.5%	n/a
	Average Reduction in Absolute Relative Error		0.0%	n/a		0.0%	n/a		-0.2%	n/a		-0.1%	n/a

Relative Errors in Predicting Employment for Late Reporters
Prior Month Employment Size <10
March 2000 - December 2002

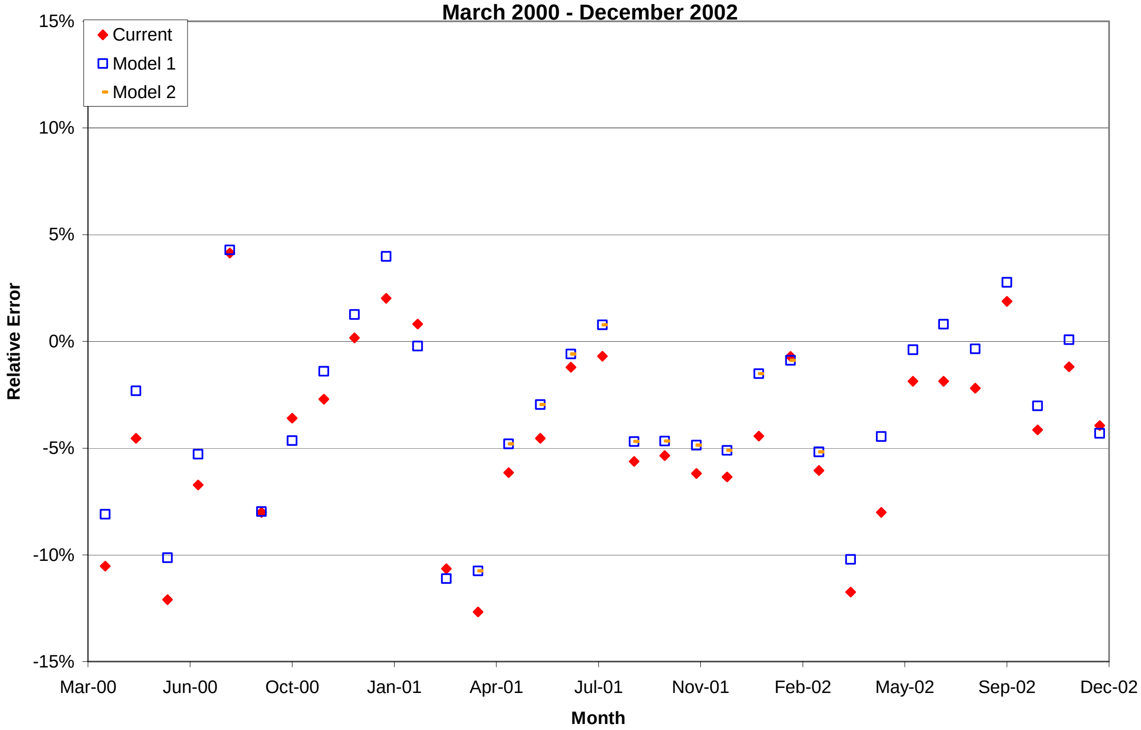
Size Class	Emp Change Class	Metric	Construction			Manufacturing			Mining			Wholesale Trade		
			Current	Model 1	Model 2	Current	Model 1	Model 2	Current	Model 1	Model 2	Current	Model 1	Model 2
<10	Low	Average Relative Error	-13.6%	-5.1%	-8.1%	-12.1%	-4.8%	-5.6%	-12.7%	-1.5%	-10.2%	-5.0%	0.4%	-8.7%
		Average Absolute Relative Error	14.2%	7.3%	8.9%	12.6%	10.1%	13.0%	20.0%	21.5%	18.2%	6.0%	4.9%	14.4%
		Average Reduction in Absolute Relative Error		6.8%	5.1%		2.6%	1.6%		-1.5%	2.9%		1.0%	-7.3%
	High	Average Relative Error	3.7%	0.9%	2.7%	-0.6%	-4.0%	-3.1%	-3.7%	-3.2%	-4.1%	2.8%	0.5%	-2.0%
		Average Absolute Relative Error	5.8%	4.7%	4.9%	6.3%	6.8%	7.5%	17.7%	22.0%	32.2%	4.6%	4.6%	16.2%
		Average Reduction in Absolute Relative Error		1.1%	-0.4%		-0.5%	0.4%		-4.4%	-9.1%		-0.1%	-14.0%
	Mid	Average Relative Error	-4.5%	-3.3%	-3.8%	-3.2%	-2.7%	-4.3%	-4.7%	-4.9%	-2.4%	-0.6%	-0.3%	-0.1%
		Average Absolute Relative Error	5.0%	4.2%	3.9%	3.9%	3.6%	4.7%	8.2%	8.6%	6.9%	1.2%	1.1%	1.0%
		Average Reduction in Absolute Relative Error		0.9%	1.1%		0.3%	0.4%		-0.4%	0.3%		0.1%	0.1%

Figure 34 presents scatterplots of the relative errors in imputed values for late reporters by month for small establishments in Construction, the industry which demonstrated the largest improvement due to Model 1. These graphs illustrate the level of improvement in predicting employment for small late reporters under Model 1. They also illustrate the aspects of imputing for larger (10+) establishments, with both current method and Model 1 subject to similar error distributions.

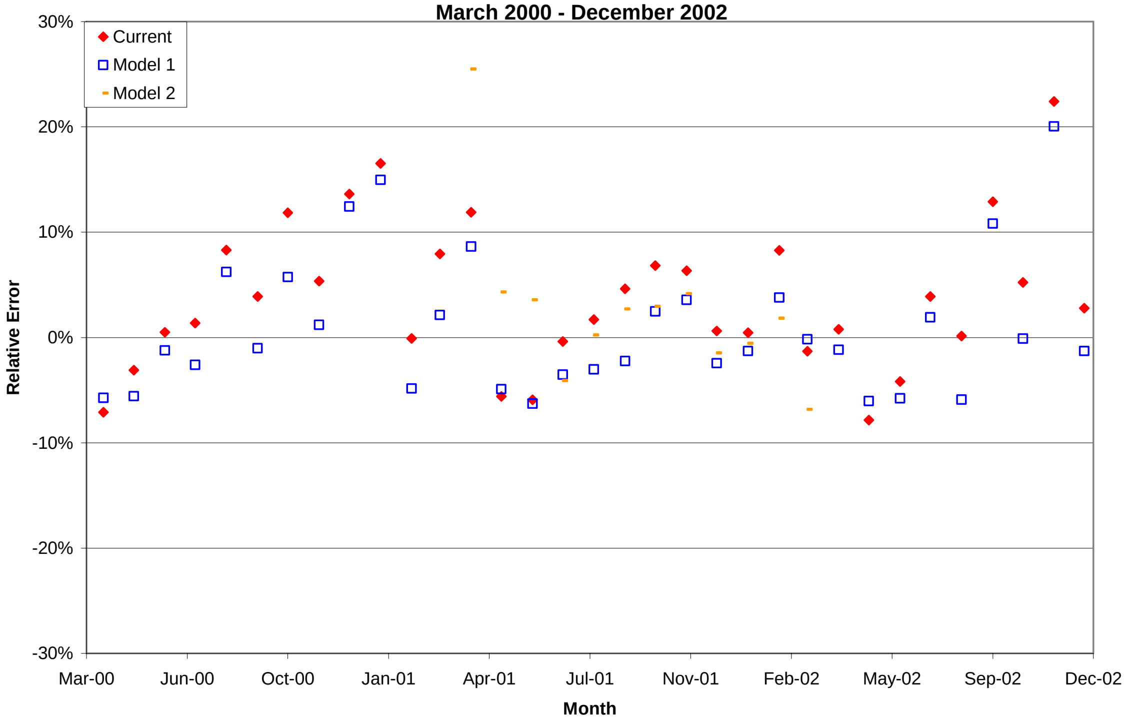
Figure 34-Relative Error in Imputed Values for Late Reporters: Construction

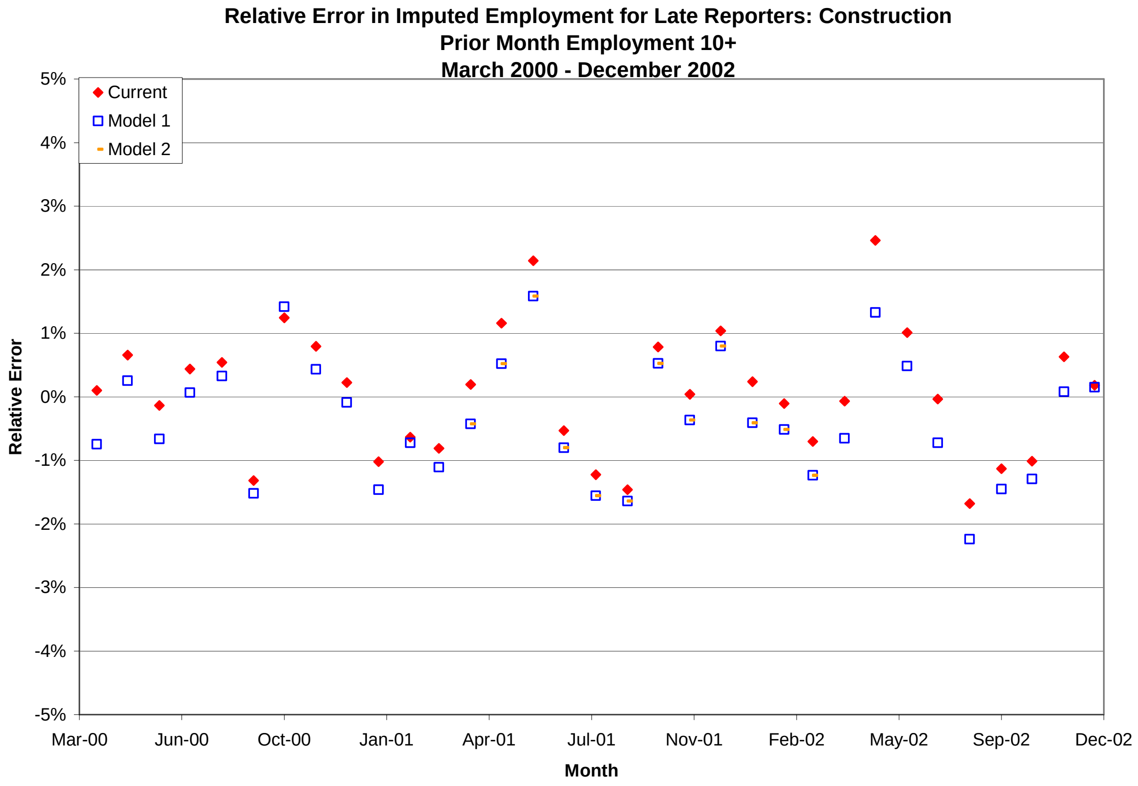


Relative Error in Imputed Employment for Late Reporters: Construction
Prior Month Employment <10, Mid Prior Month Employment Change
March 2000 - December 2002



Relative Error in Imputed Employment for Late Reporters: Construction
Prior Month Employment <10, High Prior Month Employment Change
March 2000 - December 2002





Final estimates of employment for March of 2001 and 2002 were compared to the corresponding benchmark data from ES-202, with the relative benchmark revision derived as

$$\text{Rel Rev}_{0|B}\left(\hat{Y}_{t_B}^{(1)}\right)=\frac{Y_{t_B}-\hat{Y}_t^{(2)}}{Y_{t_B}}$$

where Y_{t_B} represents the benchmark employment for month t

As seen in Table 45, benchmark revisions for Model 1A and 1B are similar to those for the current method. The differences in benchmark revisions for Model 2 are due to its being initialized with March 2001 instead of March 2002. If the current method is benchmarked to March 2001, the revisions for March 2002 are similar to those for Model 2.

Table 45-Benchmark Revisions

Benchmark Relative Revision for Estimated March Employment
March 2001, 2002

Industry	Benchmark	Current	Model 1A	Model 1B	Model 2
Construction	March 2001	1.16%	1.16%	1.31%	-
	March 2002	0.86%	0.89%	1.10%	-0.29%
Manufacturing	March 2001	1.13%	1.13%	1.14%	-
	March 2002	0.12%	0.12%	0.13%	-1.04%
Mining	March 2001	1.81%	1.76%	1.83%	-
	March 2002	0.66%	0.68%	0.87%	2.76%
Wholesale Trade	March 2001	2.10%	2.12%	2.14%	-
	March 2002	4.23%	4.27%	4.33%	2.16%

E. Summary

The current CES weighted link relative estimator is subject to bias if the expected growth rate varies by establishment characteristics within an estimation cell. Although examination of employment growth relative to prior reported information suggests the current underlying model does not hold for some subpopulations, the results obtained by imputing for missing data under the alternative models did not yield significant improvement in either monthly revisions or benchmark revisions. This is not entirely unexpected, as rough estimated biases and potential revisions were seen to be minimal. There did appear to be some support for the use of recent reported data in the working model; in particular use of such information may slightly dampen monthly revisions, especially for month-to-month change.

Given the minimal impact on both overall bias and revisions due to use of Model 1, it does not appear to afford measurable improvement over the existing model for aggregate estimates. For lower level estimates (e.g., small establishments within industry), however, Model 1 does offer the potential for improved estimates.

Chapter VI: Conclusions

Demands for timely survey estimates for economic data will continue. While methods for controlling late reporting continue to be explored and developed, this is a problem not likely to go away. In spite of efforts to improve response rates, if there is any movement, it is in the direction of higher nonresponse rates. Thus, improved methods for controlling the effects of late reporting and nonresponse will be needed.

Examination of employment growth relative to prior reported information suggests the current underlying model does not hold for some subpopulations. The resulting model misspecification has two impacts – potential differential bias for preliminary and late reporting contributing to the size of the revisions for monthly level and month-to-month change, and potential overall bias in the employment estimates contributing to the size of the benchmark revisions. The latter effect was not included in this research and warrants further investigation. This line of research should include development of approximately unbiased estimates of the $\hat{p}_{(t-1)cg}$, as well as a more in-depth examination of factors associated with large δ_{tcg} and their distributional properties. Another area of potential research is the error properties of link relative estimates resulting from alternative models.

While the model considered here attempted to utilize recently reported data for late reporters, an alternative could be considered through the use of a more direct time series approach such as that discussed by Pfeffermann and Nathan (2002).

Although the particular models selected for employment growth did not yield statistically significant improvement, there were sufficient indications of the potential for use of such parameters and approaches to warrant further research. A challenge in developing an alternative method is that the current method experiences relatively small errors in conjunction with relatively large standard errors on the link relatives. However, the sensitivity of the information is such that even very small errors can be intolerable, thus warranting consideration of methods with minimal gains.

A first approach may be to utilize a refined version of the method to monitor the results of the current method in an attempt to identify before the fact the potential for larger revisions. This could include estimation of the potential differential bias based on estimates of the $\hat{p}_{(t-1)cg}$ and δ_{tcg}, in conjunction with predicted final reporting status. Taking this approach could surface potential enhancements to the model through identification of additional factors and refinement of class definitions.

The reporting status model, on the other hand, showed very positive results. Incorporation of respondent, operational, and environmental characteristics can provide a more comprehensive accounting of the factors affecting late reporting and nonresponse. Although the focus of this research was limited to final reporting status conditional on preliminary reporting status, it is reasonable to expect such a model to perform well if used for predicting reporting status prior to data collection for the next reference period. Such a model could be used to proactively refine collection and follow up strategies.

Taking a more global view of the needs associated with a large panel survey, development and evaluation of an integrated approach to account for late reporting and nonresponse for a CES survey-type design can provide guidance as to the opportunity for error reduction in first closing estimates relative to later closings and benchmarks.

When looking at the issue of overall error of the estimates and benchmark revisions, research could also be undertaken to incorporate measurement error into the problem of estimation. Although the availability of administrative data providing actual population values is limited, it is not unusual for establishment populations. While such data are commonly used to evaluate the performance of estimators and to establish benchmarks, methods for adequately accounting for measurement error in survey estimators are lacking. Developing an understanding of the performance of resultant estimators will provide guidance to survey designers in the consideration of the use of administrative data in the estimation process.

Incorporation of the various lines of research (late reporting, nonresponse measurement error) could lead to development of an integrated approach to error adjustment for an establishment panel survey. At the least, research could result in development of a framework for generating total error estimates.

Appendices

A. Notation for a General Panel Survey

Notation and survey description will be developed first for a general panel survey, within which the CES survey fits. A more restrictive survey description will then be developed to represent the specific panel survey design to be considered in the proposed research.

1. Overview

Consider a population of fixed size N (i.e., the population does not vary over time). For each unit, $i\,(=1,...,N)$, in the population, there is a set of P variables of interest, $\mathbf{Y}_{ti[1\times P]} = \left[Y_{tip}\right]$, for each reference period $t\ \left(=1,\ldots\right)$. The set of population values across time through reference period t can be represented by the matrix

$$\mathbf{Y}_{[Nt\times P]} = \begin{bmatrix} \mathbf{Y}^T_{1[N\times P]} \\ \vdots \\ \mathbf{Y}^T_{t[N\times P]} \end{bmatrix},\ \mathbf{Y}_{t[N\times P]} = \begin{bmatrix} \mathbf{Y}_{t1[1\times P]} \\ \vdots \\ \mathbf{Y}_{tn[1\times P]} \\ \vdots \\ \mathbf{Y}_{tN[1\times P]} \end{bmatrix}$$

Statistics of interest for reference period t are the population totals for each variable, $p\left(=1,\ldots,P\right)$

$$Y_{tp} = \sum_{i=1}^{N} Y_{tip} = \begin{bmatrix} \mathbf{0}^T_N & \cdots & \mathbf{0}^T_N & \mathbf{1}^T_N \end{bmatrix}_{[1\times Nt]} \mathbf{Y}_{[Nt\times P]} \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}^T_{[P\times 1]}$$

and the change in the population totals from the prior reference period, $t-1$, to the current reference period, t, for each variable

$$\Delta_{t,(t-1)}\left(Y_p\right) = Y_{tp} - Y_{(t-1)p} =$$
$$\begin{bmatrix} \mathbf{0}_N^T & \cdots & \mathbf{0}_N^T & -\mathbf{1}_N^T & \mathbf{1}_N^T \end{bmatrix}_{[1\times Nt]} \mathbf{Y}_{[Nt\times p]} \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}_{[P\times 1]}^T$$

To obtain estimates for the statistics of interest, a panel survey is conducted, in which data, $\mathbf{y}_{ti}$, are obtained from a sample of units, $i\,(=1,...,n)$, for reference periods, $t\left(=1,\ldots\right)$. Survey estimates of the population totals, Y_{tp}, and the change in the population totals from the prior reference period, $\Delta_{t,(t-1)}\left(Y_p\right)$, are to be published soon after the reference period according to some prescribed processing schedule. The processing schedule for reference period t requires completion of data collection as of some given cutoff date, d_t, resulting in unit nonresponse. Some of the unit nonresponse is temporal, as additional responses are obtained subsequent to d_t. Given the occurrence of reporting following d_t, revised estimates for reference period t are issued as part of processing for some fixed number of subsequent reference periods. These revisions are referred to as closing estimates. The order of revision is denoted by the index variable, $k\ (=0,1,\ldots,K)$, with the original estimate referenced by $k=0$.

In addition, through some administrative data source, actual values for a subset of the variables of interest which are collected by the administrative data source, $\mathbf{Y}_{t_B}^{A}\ \left(\mathbf{Y}_{t_B} = \left[\mathbf{Y}_{t_B[N\times P_A]}^{A} \quad \mathbf{Y}_{t_B[N\times(P-P_A)]}^{A^C}\right]\right)$, for the population become available for selected reference periods, $t_B\ \left[\in\left(1,\ldots,t\right)\right]$, for which the administrative data source collects the information (referred to as benchmark reference periods), following some fixed time lag after the corresponding benchmark reference period. As a result, during survey processing for a specified reference period, survey estimates for the subset of variables of interest available from the administrative data for the most recently available benchmark reference period are replaced with the actual population values, and estimates for the remaining reference periods and for other survey variables are revised to incorporate this population information. These revisions are referred to as benchmark revisions.

Estimation for the survey involves determining how best to incorporate survey and administrative information available at the time 1st closing estimates for reference period t are processed, so as to account for nonresponse and measurement error, in addition to the sample design. One means of assessing the accuracy of the estimates is on the basis of the revisions made to incorporate late reporting and administrative data availability. There are dual objectives, those being to minimize the magnitude of revisions to estimates of Y_{tp}, the population total for the reference period, and also to minimize the magnitude of revisions to $\Delta_{t,(t-1)}\left(Y_p\right)$, the period-to-period change in the population totals.

Following is a description of the general panel survey environment. The discussion of the CES survey in Chapter III provides an illustration of the various concepts presented.

2. Survey Design

Estimates for the statistics of interest are generated using data from a panel survey, with data collected at regular intervals corresponding to the reference periods. A sample, s, of size $n\,(<N)$ is selected from the population under some probability sample design, $p\left(s\right)$. The sample design makes use of a set of Q design variables, X_{iq}, known for each unit in the population. The set of design variables can be represented by the matrix $\mathbf{X}_{[N\times Q]}=\left[X_{iq}\right]$.

Selection probabilities for the population can be represented by the vector

$$\boldsymbol{\pi}_{[N\times 1]} = \left[\pi_i\right]$$

A fixed set of sample units is surveyed every reference period. The sample selection indicator $\delta_i = 1$ indicates unit i was selected, $\delta_i = 0$ indicates unit i was not selected. The population units may be ordered such that the vector of sample selection indicators can be represented as

$$\mathbf{I}_{[N\times 1]} = \begin{bmatrix} \mathbf{1}_{[n\times 1]} \\ \mathbf{0}_{[(N-n)\times 1]} \end{bmatrix}$$

3. Data Collection

As part of data collection, sample units report values, $\mathbf{y}_{it}$. The set of sample values through reference period t can be represented by the matrix (assuming complete response)

$$\mathbf{y}_{s[nt\times P]} = \begin{bmatrix} \mathbf{y}_{s1[n\times P]} \\ \vdots \\ \mathbf{y}_{st[n\times P]} \end{bmatrix}$$

In the interest of timeliness for the publication of estimates, a cutoff date, d_t, is established for each reference period (d_t is referred to as the preliminary cutoff date for reference period t). Not all sample units report reference period t data by the preliminary cutoff date (i.e., there is nonresponse for reference period t, relative to the preliminary cutoff date). However, preliminary estimates of Y_{tp}, and $\Delta_{t,(t-1)}\left(Y_p\right)$ must be derived based upon data reported as of d_t.

Sample units not reporting reference period t data by the preliminary cutoff date may report subsequent to d_t (i.e., there is late reporting for reference period t, relative to the preliminary cutoff date). Preliminary estimates for reference period t are then revised, to incorporate late reporting, as part of survey processing for K (≥ 1) subsequent reference periods, after which time the estimates for reference period t are considered final. Data collection for reference period t thus continues through the cutoff date, d_{t+K}, which is the preliminary cutoff date for reference period $t+K$ (d_{t+K} is referred to the final cutoff date for reference period t). The estimate for reference period t generated as part of survey processing for reference period $t+k$ $\left(0 \leq k \leq K\right)$ is referred to as the k^{th} revision estimate for reference period t and is denoted as $\hat{Y}_t^{(k)}$. Thus, the preliminary estimate for reference period t is denoted as $\hat{Y}_t^{(0)}$, and the final estimate for reference period t is denoted as $\hat{Y}_t^{(K)}$.

The set of sample values for reference period t reported as of the cutoff date, d_{t+k}, is denoted as $\mathbf{y}_{st|k[n\times P]}$. The set of all sample values for all reference periods reported as of the cutoff date, d_{t+k}, can be represented by the matrix (assuming complete response)

$$\mathbf{y}_{s\bullet|k\left[n*(t+k)\times P\right]} = \left[\mathbf{y}_{s1|K\left[n\times P\right]}^T \quad \cdots \quad \mathbf{y}_{s(t-k)|K\left[n\times P\right]}^T \quad \cdots \quad \mathbf{y}_{st|k\left[n\times P\right]}^T \quad \cdots \quad \mathbf{y}_{s(t+k)|0\left[n\times P\right]}^T\right]^T$$

Correspondingly, the set of all sample values for reference period t reported as all cutoff dates, d_{t+k} $\left(0 \leq k \leq K\right)$, can be represented by the matrix (assuming complete response)

$$\mathbf{y}_{st|\bullet[n(K+1)\times P]} = \left[\mathbf{y}_{st|0[n\times P]}^T \quad \cdots \quad \mathbf{y}_{st|k[n\times P]}^T \quad \cdots \quad \mathbf{y}_{st|K[n\times P]}^T\right]^T$$

4. Response Patterns

Response indicators, $r_{it|k}$, reflect the status of reference period t data reporting for sample unit i, relative to the cutoff date, $d_{t+k}\left(0 \le k \le K\right)$. A response indicator $r_{it|k}=1$ signifies unit i reported reference period t data on or before cutoff date d_{t+k}, while a response indicator $r_{it|k}=0$ signifies unit i had not reported reference period t data as of cutoff date d_{t+k}.

Note that:

$r_{ti|k}=1 \Rightarrow r_{ti|k*}=1, \left(k \le k*\right)$

$r_{ti|k*}=r_{ti|K} \ \left(k* \ge K\right)$

Sample units may be partitioned into the following classes reflecting reference period t reporting status:

a. Preliminary Reporting $\left(PR\right)$ – unit i reported reference period t data by preliminary cutoff date

$r_{ti|0}=1$

b. Late Reporting $\left(LR\right)$ – unit i reported reference period t data after preliminary cutoff date, but on or before final cutoff date

$r_{ti|0}=0$ and $r_{ti|K}=1$

c. Nonresponse $\left(NR\right)$ – unit i did not report reference period t data as of the final cutoff date

$r_{ti|K}=0$

Response indicators for reference period t for unit i across cutoff dates may be summarized by the reporting status variable

$$\mathbf{X}_{ti} = \begin{pmatrix} X_{ti}^{PR} & X_{ti}^{LR} & X_{ti}^{NR} \end{pmatrix}^{T}$$

where the superscripts refer to preliminary reporting (PR), late reporting (LR), and

nonresponse (NR)

$$X_{ti}^{PR} = \begin{cases} 1 \text{ if } r_{ti|0} = 1 \; (\text{PR for month } t) \\ 0 \text{ if } r_{ti|0} = 0 \end{cases}$$

$$X_{ti}^{LR} = \begin{cases} 1 \text{ if } r_{ti|0} = 0 \text{ and } r_{ti|K} = 1 \; (\text{LR for month } t) \\ 0 \text{ if } r_{ti|0} = 1 \text{ or } r_{ti|K} = 0 \end{cases}$$

$$X_{ti}^{NR} = \begin{cases} 1 \text{ if } r_{ti|K} = 0 \; (\text{NR for month } t) \\ 0 \text{ if } r_{ti|K} = 1 \end{cases}$$

The set of reporting status variables for all reference periods as of cutoff date d_{t+k} can be represented by the matrix

$$\mathbf{X}_{s\bullet|k_t[n\times(t+k)]} = \begin{bmatrix} \mathbf{X}_{s1|K[n\times1]} & \cdots & \mathbf{X}_{s(t-k)|K[n\times1]} & \cdots & \mathbf{X}_{st|k[n\times1]} & \cdots & \mathbf{X}_{s(t+k)|0[n\times1]} \end{bmatrix}$$

Note that at the preliminary cutoff date for reference period t, sample units may be partitioned into only two groups relative to reference period t reporting, Preliminary Reporting $\left(\mathbf{X}_{ti} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^{T}\right)$ and Preliminary Nonreporting $\left(\mathbf{X}_{ti} = \begin{pmatrix} 1 & . & . \end{pmatrix}^{T}\right)$ (which is the aggregate of Late Reporting and Nonresponse).

5. Administrative Data

Through some administrative data source, actual values for a subset of the variables of interest, $\mathbf{Y}_{t_B[N\times P_A]}^{A}$, for the population become available for specific reference periods within each calendar year, following some fixed time lag, l. This subset of variables is referred to as benchmark data. The reference periods for which administrative data become available are designated by $t_B \left[= \left(0, (12/B), 2*(12/B), \ldots\right)\right]$, and are referred to as benchmark reference periods. As of the preliminary cutoff date for reference period t, the most recent benchmark reference period available is denoted as $t_{B|t}$, and the most recent benchmark data is denoted as $\mathbf{Y}^{A}_{t_B|t}$.

6. Estimation

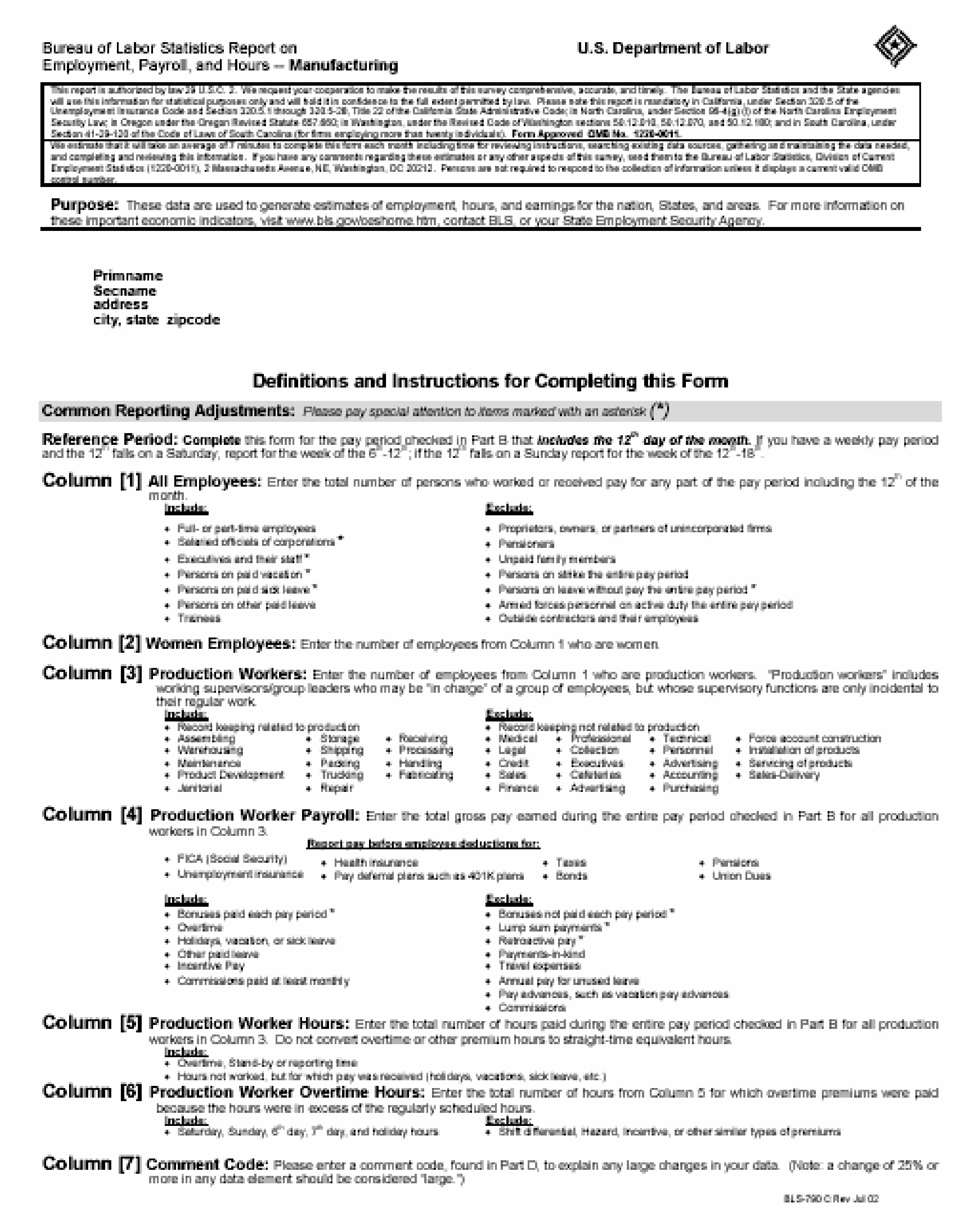
The estimates of Y_{tp}, and $\Delta_{t,(t-1)}\left(Y_p\right)$ generated as part of survey processing for reference period $t+k$ $\left(0 \le k \le K\right)$ are referred to as the k^{th} revision estimate for reference period t and are denoted by $\hat{Y}^{(k)}_{tp}$ and $\hat{\Delta}^{(k)}_{t,(t-1)}\left(Y_p\right)$. Thus, the preliminary estimates for reference period t are denoted by $\hat{Y}^{(0)}_{tp}$ and $\hat{\Delta}^{(0)}_{t,(t-1)}\left(Y_p\right)$, and the final estimates for reference period t are denoted by $\hat{Y}^{(K)}_{tp}$ and $\hat{\Delta}^{(K)}_{t,(t-1)}\left(Y_p\right)$.

The problem is how best to account for sampling, late reporting, nonresponse, and measurement error when estimating Y_{tp} and $\Delta_{t,(t-1)}\left(Y_p\right)$ based upon information available at the preliminary cutoff date for reference period t. In other words, as discussed in Chapter I, how should estimators $\hat{Y}^{(0)}_{tp}$ and $\hat{\Delta}^{(0)}_{t,(t-1)}\left(Y_p\right)$ be defined based upon the data available

$$\left[\mathbf{Y}_{B_t} : \mathbf{Z} : \mathbf{I} : \boldsymbol{\pi} : \mathbf{X}_{s\bullet|t} : \mathbf{y}_{s\bullet|t}\right]$$

B. CES Information

1. Collection Instrument – Manufacturing



Bureau of Labor Statistics Report on
Employment, Payroll, and Hours -- **Manufacturing**

U.S. Department of Labor

This report is authorized by law 29 U.S.C. 2. We request your cooperation to make the results of this survey comprehensive, accurate, and timely. The Bureau of Labor Statistics and the State agencies will use this information for statistical purposes only and will hold it in confidence to the full extent permitted by law. Please note this report is mandatory in California, under Section 320.5 of the Unemployment Insurance Code and Section 320.5.1 through 320.5-28, Title 22 of the California State Administrative Code; in North Carolina, under Section 96-4(g) (l) of the North Carolina Employment Security Law; in Oregon under the Oregon Revised Statute 657.660; in Washington, under the Revised Code of Washington sections 50.12.010, 50.12.070, and 50.12.180; and in South Carolina, under Section 41-29-120 of the Code of Laws of South Carolina (for firms employing more than twenty individuals). **Form Approved OMB No. 1220-0011.**

We estimate that it will take an average of 7 minutes to complete this form each month including time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this information. If you have any comments regarding these estimates or any other aspects of this survey, send them to the Bureau of Labor Statistics, Division of Current Employment Statistics (1220-0011), 2 Massachusetts Avenue, NE, Washington, DC 20212. Persons are not required to respond to the collection of information unless it displays a current valid OMB control number.

Purpose: These data are used to generate estimates of employment, hours, and earnings for the nation, States, and areas. For more information on these important economic indicators, visit www.bls.gov/ceshome.htm, contact BLS, or your State Employment Security Agency.

Primname
Secname
address
city, state zipcode

Definitions and Instructions for Completing this Form

Common Reporting Adjustments: *Please pay special attention to items marked with an asterisk (*)*

Reference Period: Complete this form for the pay period checked in Part B that ***includes the 12th day of the month.*** If you have a weekly pay period and the 12th falls on a Saturday, report for the week of the 6th-12th; if the 12th falls on a Sunday report for the week of the 12th-18th.

Column [1] All Employees: Enter the total number of persons who worked or received pay for any part of the pay period including the 12th of the month.

Include:
- Full- or part-time employees
- Salaried officials of corporations *
- Executives and their staff *
- Persons on paid vacation *
- Persons on paid sick leave *
- Persons on other paid leave
- Trainees

Exclude:
- Proprietors, owners, or partners of unincorporated firms
- Pensioners
- Unpaid family members
- Persons on strike the entire pay period
- Persons on leave without pay the entire pay period *
- Armed forces personnel on active duty the entire pay period
- Outside contractors and their employees

Column [2] Women Employees: Enter the number of employees from Column 1 who are women.

Column [3] Production Workers: Enter the number of employees from Column 1 who are production workers. "Production workers" includes working supervisors/group leaders who may be "in charge" of a group of employees, but whose supervisory functions are only incidental to their regular work.

Include:
- Record keeping related to production
- Assembling
- Warehousing
- Maintenance
- Product Development
- Janitorial
- Storage
- Shipping
- Packing
- Trucking
- Repair
- Receiving
- Processing
- Handling
- Fabricating

Exclude:
- Record keeping not related to production
- Medical
- Legal
- Credit
- Sales
- Finance
- Professional
- Collection
- Executives
- Cafeterias
- Advertising
- Technical
- Personnel
- Advertising
- Accounting
- Purchasing
- Force account construction
- Installation of products
- Servicing of products
- Sales-Delivery

Column [4] Production Worker Payroll: Enter the total gross pay earned during the entire pay period checked in Part B for all production workers in Column 3.

Report pay before employee deductions for:
- FICA (Social Security)
- Unemployment insurance
- Health insurance
- Pay deferral plans such as 401K plans
- Taxes
- Bonds
- Pensions
- Union Dues

Include:
- Bonuses paid each pay period *
- Overtime
- Holidays, vacation, or sick leave
- Other paid leave
- Incentive Pay
- Commissions paid at least monthly

Exclude:
- Bonuses not paid each pay period *
- Lump sum payments *
- Retroactive pay *
- Payments-in-kind
- Travel expenses
- Annual pay for unused leave
- Pay advances, such as vacation pay advances
- Commissions

Column [5] Production Worker Hours: Enter the total number of hours paid during the entire pay period checked in Part B for all production workers in Column 3. Do not convert overtime or other premium hours to straight-time equivalent hours.

Include:
- Overtime, Stand-by or reporting time
- Hours not worked, but for which pay was received (holidays, vacations, sick leave, etc.)

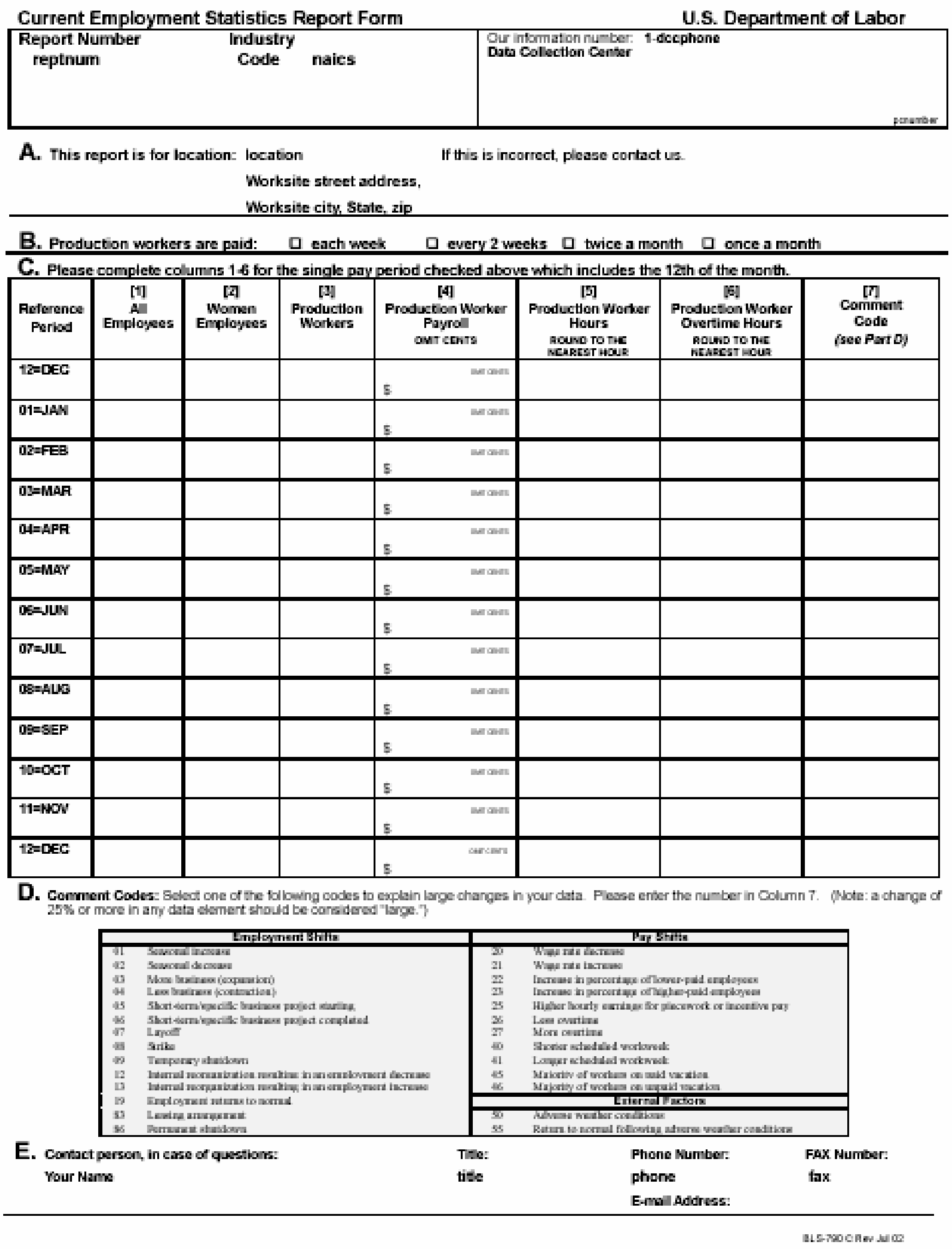
Column [6] Production Worker Overtime Hours: Enter the total number of hours from Column 5 for which overtime premiums were paid because the hours were in excess of the regularly scheduled hours.

Include:
- Saturday, Sunday, 6th day, 7th day, and holiday hours

Exclude:
- Shift differential, Hazard, Incentive, or other similar types of premiums

Column [7] Comment Code: Please enter a comment code, found in Part D, to explain any large changes in your data. (Note: a change of 25% or more in any data element should be considered "large.")

BLS-790 C Rev Jul 02



Current Employment Statistics Report Form

U.S. Department of Labor

Report Number reptnum	Industry Code naics	Our information number: 1-dccphone Data Collection Center pcnumber

A. This report is for location: location
Worksite street address,
Worksite city, State, zip

If this is incorrect, please contact us.

B. Production workers are paid: □ each week □ every 2 weeks □ twice a month □ once a month

C. Please complete columns 1-6 for the single pay period checked above which includes the 12th of the month.

Reference Period	[1] All Employees	[2] Women Employees	[3] Production Workers	[4] Production Worker Payroll OMIT CENTS	[5] Production Worker Hours ROUND TO THE NEAREST HOUR	[6] Production Worker Overtime Hours ROUND TO THE NEAREST HOUR	[7] Comment Code *(see Part D)*
12=DEC				OMIT CENTS $			
01=JAN				OMIT CENTS $			
02=FEB				OMIT CENTS $			
03=MAR				OMIT CENTS $			
04=APR				OMIT CENTS $			
05=MAY				OMIT CENTS $			
06=JUN				OMIT CENTS $			
07=JUL				OMIT CENTS $			
08=AUG				OMIT CENTS $			
09=SEP				OMIT CENTS $			
10=OCT				OMIT CENTS $			
11=NOV				OMIT CENTS $			
12=DEC				OMIT CENTS $			

D. **Comment Codes:** Select one of the following codes to explain large changes in your data. Please enter the number in Column 7. (Note: a change of 25% or more in any data element should be considered "large.")

Employment Shifts		Pay Shifts	
01	Seasonal increase	20	Wage rate decrease
02	Seasonal decrease	21	Wage rate increase
03	More business (expansion)	22	Increase in percentage of lower-paid employees
04	Less business (contraction)	23	Increase in percentage of higher-paid employees
05	Short-term/specific business project starting	25	Higher hourly earnings for piecework or incentive pay
06	Short-term/specific business project completed	26	Less overtime
07	Layoff	27	More overtime
08	Strike	40	Shorter scheduled workweek
09	Temporary shutdown	41	Longer scheduled workweek
12	Internal reorganization resulting in an employment decrease	45	Majority of workers on paid vacation
13	Internal reorganization resulting in an employment increase	46	Majority of workers on unpaid vacation
19	Employment returns to normal	External Factors	
83	Leasing arrangement	50	Adverse weather conditions
86	Permanent shutdown	55	Return to normal following adverse weather conditions

E. Contact person, in case of questions:
Your Name

Title:
title

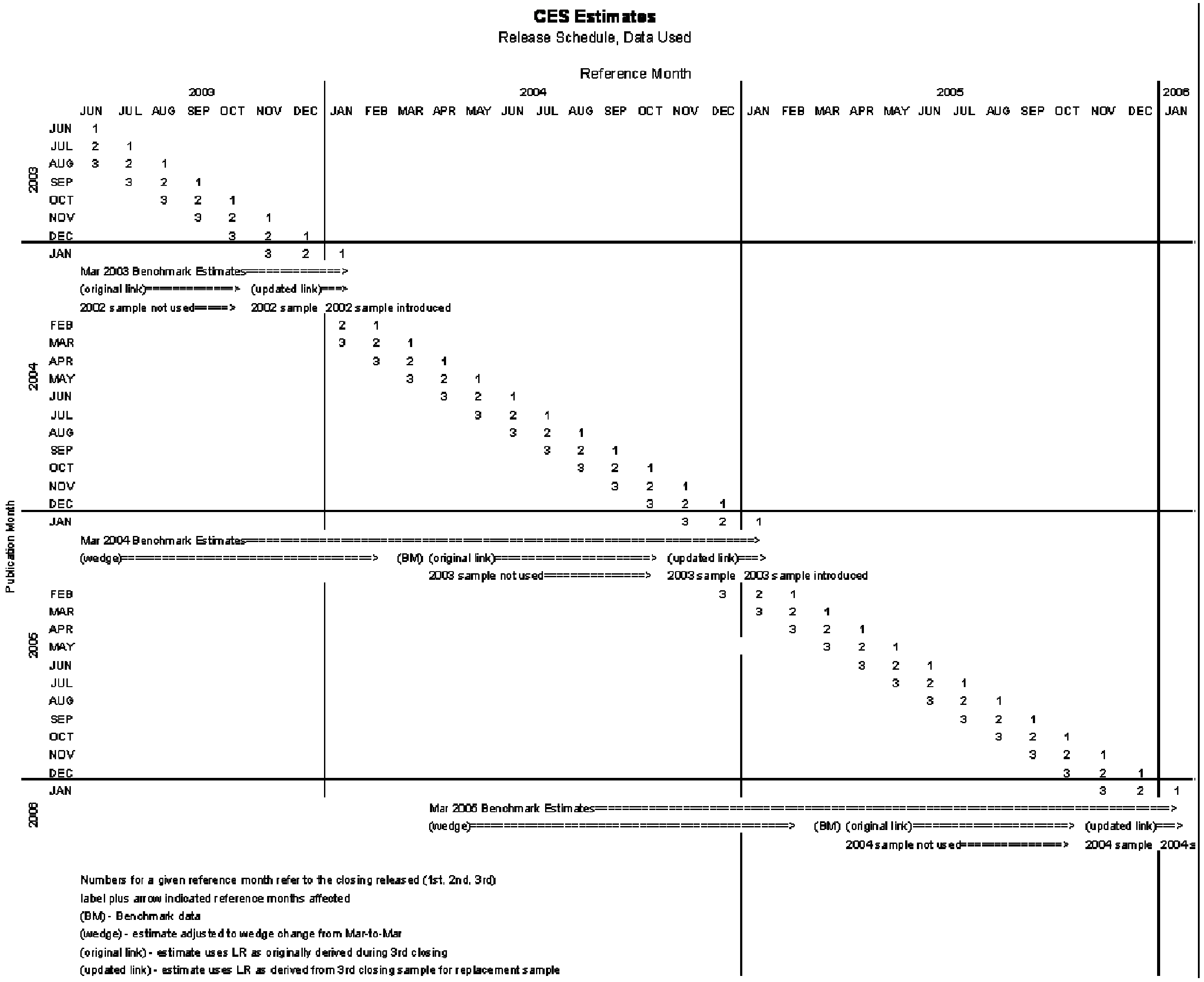
Phone Number:
phone

FAX Number:
fax

E-mail Address:

BLS-790C Rev Jul 02

2. CES Estimate Revision Schedule



CES Estimates

Release Schedule, Data Used

Reference Month

		2003							2004												2005												2006
Publication Month		JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN
2003	JUN	1																															
	JUL	2	1																														
	AUG	3	2	1																													
	SEP		3	2	1																												
	OCT			3	2	1																											
	NOV				3	2	1																										
	DEC					3	2	1																									
2004	JAN						3	2	1																								
		Mar 2003 Benchmark Estimates============>																															
		(original link)============> (updated link)===>																															
		2002 sample not used====> 2002 sample 2002 sample introduced																															
	FEB								2	1																							
	MAR								3	2	1																						
	APR									3	2	1																					
	MAY										3	2	1																				
	JUN											3	2	1																			
	JUL												3	2	1																		
	AUG													3	2	1																	
	SEP														3	2	1																
	OCT															3	2	1															
	NOV																3	2	1														
	DEC																	3	2	1													
2005	JAN																		3	2	1												
		Mar 2004 Benchmark Estimates==>																															
		(wedge)==========================> (BM) (original link)==================> (updated link)===>																															
		2003 sample not used=============> 2003 sample 2003 sample introduced																															
	FEB																			3	2	1											
	MAR																				3	2	1										
	APR																					3	2	1									
	MAY																						3	2	1								
	JUN																							3	2	1							
	JUL																								3	2	1						
	AUG																									3	2	1					
	SEP																										3	2	1				
	OCT																											3	2	1			
	NOV																												3	2	1		
	DEC																													3	2	1	
2006	JAN																														3	2	1
		Mar 2005 Benchmark Estimates==>																															
		(wedge)====================================> (BM) (original link)==================> (updated link)===>																															
		2004 sample not used=============> 2004 sample 2004s																															

Numbers for a given reference month refer to the closing released (1st, 2nd, 3rd)
label plus arrow indicated reference months affected
(BM) - Benchmark data
(wedge) - estimate adjusted to wedge change from Mar-to-Mar
(original link) - estimate uses LR as originally derived during 3rd closing
(updated link) - estimate uses LR as derived from 3rd closing sample for replacement sample

C. Selected Program Code for Data Preparation

1. Reading CES Microdata

```
Filename ces00 "c:\CES Data\micro.y2000.sam0001.txt";
Filename ces01 "c:\CES Data\micro.y2001.sam0001.txt";
Filename ces0203 "c:\CES Data\micro.y2002.y2003.sam0001.txt7";
libname hold "c:\CES Data";

*Program Name: x:/Research Project/File Creation/Edited Microdata Read
*This program reads the CES microdata files,
*runs an additional edit to look for anomalous changes,
*creates monthly response indicators:
*NR = 1 - not reported (by 3rd closing)
*LR = 1 - late reporter (2nd or 3rd closing)
*and outputs a file with data for 2000-2002;

*read CES microdata for 2000;

%macro out1(mon);
data
%do a=1 %to &mon;
        ces_00_&a (keep=ces_id LR&a NR&a ae&a atyp&a flag&a)
%end;
 check_00 (keep=ces_id sam00 sam01) close_00 (keep=ces_id close);
%mend;

%macro recode1(mon);
%do a=1 %to &mon;
        if month=&a then do;
                LR&a=0;
                NR&a=0;
                ae&a=ae;
                atyp&a=0;
                flag&a=0;

                *change NR indicator to 1 if ae lt 0;

                if ae lt 0 then do;
                        NR&a=1;
                        ae&a=.;
                        flag&a=1;
                end;

                *change NR indicator to 1 if ae >99,999;
```

```
else if ae gt 99999 then do;
          NR&a=1;
          ae&a=.;
          flag&a=2;
end;

*change NR indicator to 1 if ae missing;

else if ae = . then do;
          NR&a=1;
          flag&a=3;
end;

*change NR indicator to 1 if close missing;

else if close=. then do;
          NR&a=1;
          flag&a=4;
end;

*change NR indicator to 1 if close gt 3;
*(happens later in process,;
*after all months are merged;
*so ae figure can be used in calculating ave_ae);
*check for close > 4;

else if close gt 3 then  do;
          flag&a=5;
          if close >4 then output close_00;
end;

*change LR indicator to 1 if NR = 0 and close = (2 or 3);

else if close gt 1 then LR&a=1;

*set atyp to 2 if explan=90;
*all data unusable for that month;
*if not unusable next month,;
*this month can be used for next months LR;

if explan = 90 then atyp&a=2;

*set atyp to 1 if class is an odd # and cc ne 90;
*ae data treated as unweighted for that months LR;
*if not atypical next month,;
```

```
        *this month can be used for next months LR;

        else if class gt 0 then do;
            k1=class/2;
            k2=(class-1)/2;
            if floor(k1)=floor(k2) then     do;
                atyp&a=1;
            end;
        end;
        output ces_00_&a;
    end;
%end;
%mend;

%macro sort1(mon);
%do a=1 %to &mon;
    proc sort data=ces_00_&a;
    by ces_id;
    run;
%end;
%mend;

%macro merge1(mon);
merge
%do a=1 %to &mon;
    ces_00_&a
%end;
;
%mend;


%out1(12);
infile ces00 missover;
input @1 month 2.
        @3 year 4.
   @7 ces_id 9.
   @20 ae 6.
        @58 close 1.
        @60 explan 2.
        @62 class 2.
        @64 sam00 1.
        @65 sam01 1.
        @;
if sam00=1 then do;
%recode1(12);
end;
```

```
else output check_00;
run;

*Look at records not pulled into ces_00 files to make sure sam00 not odd;

proc sort data=check_00;
by sam00;
run;

proc freq data=check_00;
tables sam00*sam01;
run;

*Look at records pulled into ces_00 files with unexpected close;

proc sort data=close_00;
by close;
run;

proc freq data=check_00;
tables close;
run;

*Process the ces_00 files;

%sort1(12);
run;

data ces_00;
%merge1(12);
by ces_id;
run;

*read CES microdata for 2001;

%macro out2(mon);
data
%do a=13 %to &mon;
        ces_01_&a (keep=ces_id LR&a NR&a ae&a atyp&a flag&a)
%end;
  check_01 (keep=ces_id sam00 sam01) close_01 (keep=ces_id close);
%mend;

%macro recode2(mon);
%do a=13 %to &mon;
        if month=&a-12 then do;
```

```
LR&a=0;
NR&a=0;
ae&a=ae;
atyp&a=0;
flag&a=0;

*change NR indicator to 1 if ae lt 0;

if ae lt 0 then do;
          NR&a=1;
          ae&a=.;
          flag&a=1;
end;

*change NR indicator to 1 if ae >99,999;

else if ae gt 99999 then do;
          NR&a=1;
          ae&a=.;
          flag&a=2;
end;

*change NR indicator to 1 if ae missing;

else if ae = . then do;
          NR&a=1;
          flag&a=3;
end;

*change NR indicator to 1 if close missing;

else if close=. then do;
          NR&a=1;
          flag&a=4;
end;

*change NR indicator to 1 if close gt 3;
*(happens later in process,;
*after all months are merged;
*so ae figure can be used in calculating ave_ae);
*check for close > 4;

else if close gt 3 then  do;
          flag&a=5;
          if close >4 then output close_01;
end;
```

```
        *change LR indicator to 1 if NR = 0 and close = (2 or 3);

        else if close gt 1 then LR&a=1;

        *set atyp to 2 if explan=90;
        *all data unusable for that month;
        *if not unusable next month,;
        *this month can be used for next months LR;

        if explan = 90 then atyp&a=2;

        *set atyp to 1 if class is an odd # and cc ne 90;
        *ae data treated as unweighted for that months LR;
        *if not atypical next month,;
        *this month can be used for next months LR;

        else if class gt 0 then do;
            k1=class/2;
            k2=(class-1)/2;
            if floor(k1)=floor(k2) then    do;
                atyp&a=1;
            end;
        end;
        output ces_01_&a;
    end;
%end;
%mend;

%macro sort2(mon);
%do a=13 %to &mon;
    proc sort data=ces_01_&a;
    by ces_id;
    run;
%end;
%mend;

%macro merge2(mon);
merge
%do a=13 %to &mon;
    ces_01_&a
%end;
;
%mend;
```

```
%out2(24);
infile ces01 missover;
input @1 month 2.
                        @3 year 4.
        @7 ces_id 9.
        @20 ae 6.
                        @58 close 1.
                        @60 explan 2.
                        @62 class 2.
                        @64 sam00 1.
                        @65 sam01 1.
                        @;
if sam00=1 then do;
%recode2(24);
end;
else output check_01;
run;

*Look at records not pulled into ces_01 files to make sure sam00 not odd;

proc sort data=check_01;
by sam00;
run;

proc freq data=check_01;
tables sam00*sam01;
run;

*Look at records pulled into ces_01 files with unexpected close;

proc sort data=close_01;
by close;
run;

proc freq data=check_01;
tables close;
run;

*Process the ces_01 files;

%sort2(24);
run;

data ces_01;
%merge2(24);
by ces_id;
```

```
run;

*read CES microdata for 2002/2003;

%macro out3(mon);
data
%do a=25 %to &mon;
        ces_02_&a (keep=ces_id LR&a NR&a ae&a atyp&a flag&a)
%end;
        check_02 (keep=ces_id sam00 sam01) close_02 (keep=ces_id close)
                check_03 (keep=ces_id month year sam00 sam01);
%mend;

%macro recode3(mon);
%do a=25 %to &mon;
        if month=&a-24 then do;
                LR&a=0;
                NR&a=0;
                ae&a=ae;
                atyp&a=0;
                flag&a=0;

                *change NR indicator to 1 if ae lt 0;

                if ae lt 0 then do;
                        NR&a=1;
                        ae&a=.;
                        flag&a=1;
                end;

                *change NR indicator to 1 if ae >99,999;

                else if ae gt 99999 then do;
                        NR&a=1;
                        ae&a=.;
                        flag&a=2;
                end;

                *change NR indicator to 1 if ae missing;

                else if ae = . then do;
                        NR&a=1;
                        flag&a=3;
                end;

                *change NR indicator to 1 if close missing;
                else if close=. then do;
                        NR&a=1;
                        flag&a=4;
                end;

                *change NR indicator to 1 if close gt 3;
                *(happens later in process,;
                *after all months are merged;
                *so ae figure can be used in calculating ave_ae);
                *check for close > 4;

                else if close gt 3 then  do;
                        flag&a=5;
                        if close >4 then output close_02;
                end;

                *change LR indicator to 1 if NR = 0 and close = (2 or 3);

                else if close gt 1 then LR&a=1;

                *set atyp to 2 if explan=90;
                *all data unusable for that month;
                *if not unusable next month,;
                *this month can be used for next months LR;

                if explan = 90 then atyp&a=2;

                *set atyp to 1 if class is an odd # and cc ne 90;
                *ae data treated as unweighted for that months LR;
                *if not atypical next month,;
                *this month can be used for next months LR;

                else if class gt 0 then do;
                        k1=class/2;
                        k2=(class-1)/2;
                        if floor(k1)=floor(k2) then      do;
                            atyp&a=1;
                        end;
                end;
                output ces_02_&a;
        end;
%end;
%mend;

%macro sort3(mon);
```

```
%do a=25 %to &mon;
        proc sort data=ces_02_&a;
        by ces_id;
        run;
%end;
%mend;

%macro merge3(mon);
merge
%do a=25 %to &mon;
        ces_02_&a
%end;
;
%mend;


%out3(36);
infile ces0203 missover;
input @1 month 2.
                @3 year 4.
      @7 ces_id 9.
      @20 ae 6.
                @58 close 1.
                @60 explan 2.
                @62 class 2.
                @64 sam00 1.
                @65 sam01 1.
                @;
if sam00=1 then do;
if year=2002 then do;
%recode3(36);
end;
else if year=2003 then output check_03;
end;
else output check_02;
run;

*Look at records not pulled into ces_02 files to make sure sam00 not odd;

proc sort data=check_02;
by sam00;
run;

proc freq data=check_02;
tables sam00*sam01;
run;
```

```
*Look at records pulled into ces_02 files with unexpected close;

proc sort data=close_02;
by close;
run;

proc freq data=check_02;
tables close;
run;

*Process the ces_02 files;

%sort3(36);
run;

data ces_02;
%merge3(36);
by ces_id;
run;

*merge data for 2000, 2001, 2002;

%macro out4(mon);
data editces (drop=ae0 NR0
%do a=1 %to &mon;
        edae&a
%end;
);
%mend;

%macro recode4(mon);
%do a=1 %to &mon;
        if NR&a=. then do;
                NR&a=1;
                LR&a=0;
                atyp&a=0;
                flag&a=6;
        end;
%end;
%mend;

*determine the first and last month reported;
*first month requires a response within 1st-3rd closing,;
*but accepts edit failures and atypicals;
*last month merely requires a response;
```

```
%macro firstlast1(mon);
first_mo=0;
last_mo=0;
%do a=1 %to &mon;
        if first_mo=0 then do;
                if NR&a=0 then first_mo=&a;
        end;
        if NR&a=0 then last_mo=&a;
%end;
%mend;

*conduct custom edit;
*flag as atypical if month-to-month change is > 100 and;
*month-to-month change is > 1.5 times average ae for two months;

%macro clean1(mon);
ae0=.;
NR0=1;
%do a=1 %to &mon;
        %do b=&a-1 %to &a-1;
                if NR&a=0 and NR&b=0 and atyp&a=0 then do;
                        if abs(ae&a-ae&b) gt 100 then do;
                                if abs((ae&a-ae&b)/(.5*(ae&a+ae&b))) gt 1.5 then
atyp&a=3;
                        end;
                end;
        %end;
        if flag&a=5 then NR&a=1;
%end;
%mend;

*calculate an edited ae by deleting ae if atyp > 0 and;
* next months atyp > 0 or missing;

%macro clean2(mon);
%do a=1 %to &mon;
        %do b=&a+1 %to &a+1;
                if atyp&a gt 0 then do;
                        if atyp&b gt 0 then edae&a=.;
                        else if atyp&b=. then edae&a=.;
                        else edae&a=ae&a;
                end;
        %end;
%end;
%mend;
```

```
%out4(36);
merge ces_00 ces_01 ces_02;
by ces_id;
%recode4(36);
%clean1(36);
%clean2(35);
if atyp36 gt 0 then edae36=.;
else edae36=ae36;
%firstlast1(36);

*calculate average employment based on;
*reported months (regardless of close),;
*reported months with no atyp or atyp followed by non-atyp;
*for use in weighting counts;

ed_ave_ae=mean(edae1,edae2,edae3,edae4,edae5,edae6,
          edae7,edae8,edae9,edae10,edae11,edae12,
          edae13,edae14,edae15,edae16,edae17,edae18,
          edae19,edae20,edae21,edae22,edae23,edae24,
          edae25,edae26,edae27,edae28,edae29,edae30,
          edae31,edae32,edae33,edae34,edae35,edae36);
ave_ae=mean(ae1,ae2,ae3,ae4,ae5,ae6,
          ae7,ae8,ae9,ae10,ae11,ae12,
          ae13,ae14,ae15,ae16,ae17,ae18,
          ae19,ae20,ae21,ae22,ae23,ae24,
          ae25,ae26,ae27,ae28,ae29,ae30,
          ae31,ae32,ae33,ae34,ae35,ae36);
diff_ae=ave_ae-ed_ave_ae;
run;

*delete records with no first month report;

data editces nofirst;
set editces;
if first_mo=0 then output nofirst;
else output editces;
run;

*Look at records deleted due to no first month;

proc sort data=nofirst;
by last_mo;
run;

proc freq data=nofirst;
```

```
tables last_mo;
run;

proc univariate data=nofirst;
var ave_ae ed_ave_ae diff_ae;
run;

*Look at characteristics of editces;

proc freq data=editces;
tables first_mo last_mo
        NR1 NR2 NR3 NR4 NR5 NR6 NR7 NR8 NR9 NR10
        NR11 NR12 NR13 NR14 NR15 NR16 NR17 NR18 NR19 NR20
        NR21 NR22 NR23 NR24 NR25 NR26 NR27 NR28 NR29 NR30
        NR31 NR32 NR33 NR34 NR35 NR36
        LR1 LR2 LR3 LR4 LR5 LR6 LR7 LR8 LR9 LR10
        LR11 LR12 LR13 LR14 LR15 LR16 LR17 LR18 LR19 LR20
        LR21 LR22 LR23 LR24 LR25 LR26 LR27 LR28 LR29 LR30
        LR31 LR32 LR33 LR34 LR35 LR36
        atyp1 atyp2 atyp3 atyp4 atyp5 atyp6 atyp7 atyp8 atyp9 atyp10
        atyp11 atyp12 atyp13 atyp14 atyp15 atyp16 atyp17 atyp18 atyp19 atyp20
        atyp21 atyp22 atyp23 atyp24 atyp25 atyp26 atyp27 atyp28 atyp29 atyp30
        atyp31 atyp32 atyp33 atyp34 atyp35 atyp36
        flag1 flag2 flag3 flag4 flag5 flag6 flag7 flag8 flag9 flag10
        flag11 flag12 flag13 flag14 flag15 flag16 flag17 flag18 flag19 flag20
        flag21 flag22 flag23 flag24 flag25 flag26 flag27 flag28 flag29 flag30
        flag31 flag32 flag33 flag34 flag35 flag36;
run;

proc univariate data=editces;
var ave_ae ed_ave_ae diff_ae
        ae1 ae2 ae3 ae4 ae5 ae6 ae7 ae8 ae9 ae10
        ae11 ae12 ae13 ae14 ae15 ae16 ae17 ae18 ae19 ae20
        ae21 ae22 ae23 ae24 ae25 ae26 ae27 ae28 ae29 ae30
        ae31 ae32 ae33 ae34 ae35 ae36;
run;

data hold.editces (drop=diff_ae);
set editces;
run;

proc contents data=hold.editces;
run;
```

2. Obtaining NAICS from CES cross-walk file

```
libname hold "c:\CES Data";

*Program Name: x:Research Project/File Creation/CW_NAICS
*This program creates NAICS groupings for CW file;


data hold.cw_mar_03 (keep=ldbnum ldbae naics_00 naics_01 naics_cw
          naics_sec report reptwith run_00 run_01 ui_00 ui_01 sam_00 sam_01);
set hold.cw_10mar03;
if naics_00 ge 900000 then naics_cw="govt";
else if naics_00 ge 800000 then naics_cw="othsvcs";
else if naics_00 ge 700000 then naics_cw="leisure";
else if naics_00 ge 600000 then naics_cw="educ";
else if naics_00 ge 540000 then naics_cw="prof";
else if naics_00 ge 530000 then naics_cw="fire";
else if naics_00 ge 520000 then naics_cw="fire";
else if naics_00 ge 510000 then naics_cw="info";
else if naics_00 ge 480000 then naics_cw="tpu";
else if naics_00 ge 440000 then naics_cw="retail";
else if naics_00 ge 420000 then naics_cw="whole";
else if naics_00 ge 310000 then naics_cw="mfg";
else if naics_00 ge 230000 then naics_cw="construct";
else if naics_00 ge 220000 then naics_cw="tpu";
else if naics_00 ge 210000 then naics_cw="mining";
else if naics_00 ge 114000 then naics_cw="agr";
else if naics_00 ge 113300 then naics_cw="mining";
else if naics_00 ge 111000 then naics_cw="agr";
else naics_cw="miss";
naics_sec=naics_cw;
if naics_sec="othsvcs" then naics_sec="svcs";
else if naics_sec="leisure" then naics_sec="svcs";
else if naics_sec="educ" then naics_sec="svcs";
else if naics_sec="prof" then naics_sec="svcs";
else if naics_sec="info" then naics_sec="svcs";
run;

proc contents;
run;
```

3. Merging CES microdata and CES cross-walk files

```
libname hold "c:\CES Data";

*Program Name: x:Research Project/File Creation/merge_ces_cw;
*This program merges the 2000 CES sample records with
*the 3/10/03 CW file records.
*Merging is based on ces_id (ces) to report (cw),
*first to parent records, then (if unmatached) to child records.
*The only data from ces is ces_id other variables are kept from cw
*An output data set, ces_cw, is created,
*with added field source (1 - ces & cw parent,
*2 - ces & cw child, 3 - ces only);

data ces_full;
set hold.editces(keep=ces_id);
rpt=ces_id;
run;

proc sort;
by rpt;
run;

data nochild parent child;
set hold.cw_mar_03 (keep=report reptwith ldbnum ldbae
        ui_00 run_00 naics_cw sam_00 sam_01);
if sam_00=1 then do;
        rptw=10;
        rptw=reptwith;
        rpt=10;
        rpt=report;
        if reptwith=. then output nochild;
        else if reptwith=report then output parent;
        else output child;
end;
run;

proc sort data=parent;
by rpt;
run;

proc sort data=child;
by rpt;
run;
```

```
data ces_p only_ces_p only_p;
merge ces_full(in=a) parent(in=b);
by rpt;
if a & b then do;
          source=1;
          output ces_p;
end;
else if a then output only_ces_p;
else if b then output only_p;
run;

data ces_c only_ces_c only_c;
merge only_ces_p(in=a) child(in=b);
by rpt;
if a & b then do;
          source=2;
          output ces_c;
end;
else if a then do;
          source=3;
          output only_ces_c;
end;
else if b then output only_c;
run;

data ces_cw;
set ces_p ces_c only_ces_c;
run;

proc sort data=ces_cw;
by ces_id;
run;

data hold.ces_cw;
set ces_cw;
run;

data parent;
set only_p;
run;

data child;
set only_c;
run;

proc freq data=hold.ces_cw;
```

```
tables naics_cw;
run;

proc freq data=parent;
tables naics_cw;
run;

proc freq data=child;
tables naics_cw;
run;

proc contents data=hold.ces_cw;
run;
```

4. Appending NAICS onto CES microdata

```
libname hold "c:\CES Data";

*Program Name: x:Research Project/File Creation/Edited CES_NAICS
*This program appends the NAICS code from CW file
*to the Edited CES data file - the resulting file is sorted by NAICS;

data naics (drop=ui_00 run_00);
set hold.ces_cw (keep=ces_id naics_cw ui_00 run_00 ldbnum rpt rptw source);
ui=ui_00;
run=run_00;
run;

proc sort data=hold.editces out=editces;
by ces_id;
run;

data ces_naics;
merge editces (in=a) naics (in=b);
by ces_id;
if a;
if naics_cw='    ' then delete;
else if naics_cw='agr ' then delete;
else if naics_cw='govt' then delete;
else if naics_cw='miss' then delete;
else if naics_cw='oths' then delete;
else if naics_cw='educ' then delete;
else if naics_cw='prof' then delete;
else if naics_cw='info' then delete;
else if naics_cw='leis' then delete;
else if naics_cw='reta' then delete;
else if naics_cw='tpu ' then delete;
else if naics_cw='fire' then delete;
run;

proc freq data=ces_naics;
tables first_mo last_mo naics_cw
        NR1 NR2 NR3 NR4 NR5 NR6 NR7 NR8 NR9 NR10
        NR11 NR12 NR13 NR14 NR15 NR16 NR17 NR18 NR19 NR20
        NR21 NR22 NR23 NR24 NR25 NR26 NR27 NR28 NR29 NR30
        NR31 NR32 NR33 NR34 NR35 NR36
        LR1 LR2 LR3 LR4 LR5 LR6 LR7 LR8 LR9 LR10
        LR11 LR12 LR13 LR14 LR15 LR16 LR17 LR18 LR19 LR20
        LR21 LR22 LR23 LR24 LR25 LR26 LR27 LR28 LR29 LR30
        LR31 LR32 LR33 LR34 LR35 LR36
        atyp1 atyp2 atyp3 atyp4 atyp5 atyp6 atyp7 atyp8 atyp9 atyp10
        atyp11 atyp12 atyp13 atyp14 atyp15 atyp16 atyp17 atyp18 atyp19 atyp20
        atyp21 atyp22 atyp23 atyp24 atyp25 atyp26 atyp27 atyp28 atyp29 atyp30
        atyp31 atyp32 atyp33 atyp34 atyp35 atyp36
        flag1 flag2 flag3 flag4 flag5 flag6 flag7 flag8 flag9 flag10
        flag11 flag12 flag13 flag14 flag15 flag16 flag17 flag18 flag19 flag20
        flag21 flag22 flag23 flag24 flag25 flag26 flag27 flag28 flag29 flag30
        flag31 flag32 flag33 flag34 flag35 flag36
        naics_cw*LR1*NR1 naics_cw*atyp1
        naics_cw*LR12*NR12 naics_cw*atyp12
        naics_cw*LR24*NR24 naics_cw*atyp24
        naics_cw*LR36*NR36 naics_cw*atyp36;
run;

proc univariate data=ces_naics;
var ave_ae ed_ave_ae ae1 ae2 ae3 ae4 ae5 ae6 ae7 ae8 ae9 ae10
        ae11 ae12 ae13 ae14 ae15 ae16 ae17 ae18 ae19 ae20
        ae21 ae22 ae23 ae24 ae25 ae26 ae27 ae28 ae29 ae30
        ae31 ae32 ae33 ae34 ae35 ae36;
run;

proc sort data=ces_naics out=hold.editces1;
by ui;
run;

proc contents data=hold.editces1;
run;
```

5. Appending length of pay period from August 2001 CES registry file

```
libname hold "c:\CES Data";

*Program Name: x:/Research Project/File Creation/Registry File
*;


data aug01 (drop=rptid);
set hold.aug01 (keep=ui rptid rptw lopp respcode);
rpt=rptid;
run;

proc sort data=aug01;
by rpt;
run;

data aug01_1 aug01_2;
set aug01;
by rpt;
if first.rpt then output aug01_1;
else output aug01_2;
run;

data ces_cw (drop=ui_00);
set hold.ces_cw (keep=ces_id ui_00 run_00 naics_cw rpt rptw);
if naics_cw='oths' then delete;
else if naics_cw='agr ' then delete;
else if naics_cw='miss' then delete;
else if naics_cw='govt' then delete;
else if naics_cw='educ' then delete;
else if naics_cw='prof' then delete;
else if naics_cw='info' then delete;
else if naics_cw='leis' then delete;
else if naics_cw='reta' then delete;
else if naics_cw='tpu ' then delete;
else if naics_cw='fire' then delete;
ui=0;
ui=ui_00;
run;

proc sort data=ces_cw;
by rpt;
run;
```

```
data ces_cw1 ces_cw2;
set ces_cw;
by rpt;
if first.rpt then output ces_cw1;
else output ces_cw2;
run;

proc sort data=aug01;
by rpt;
run;

data aug01_1 aug01_2;
set aug01;
by rpt;
if first.rpt then output aug01_1;
else output aug01_2;
run;

data ces1_aug1 ces1_only1;
merge ces_cw1 (in=a)
        aug01_1 (in=b);
by rpt;
if a & b then output ces1_aug1;
else if a then output ces1_only1;
run;

proc sort data=ces1_aug1;
by ces_id;
run;

data ces1_aug11 ces1_aug2;
set ces1_aug1;
by ces_id;
if first.ces_id then output ces1_aug11;
else output ces1_aug2;
run;

proc sort data= hold.editces1 out=ces;
by ces_id;
run;

data hold.editces2;
merge ces (in=a) ces1_aug11 (in=b);
by ces_id;
if a then output hold.editces2;
run;
```

6. Appending sample design information from CES random group file

```
libname hold "c:\CES Data";

*Program Name: x:Research Project/File Creation/Edited CES_NAICS_RG
*This program appends size, state, selection weight, and RG values from
RANGROUP file
*to the Edited CES data file - the resulting file is sorted by ui;

data rangroup;
set hold.rangroup (keep=ui selwt size st h1-h80);
run;

proc sort data=rangroup nodupkey;
by ui;
run;

data editces1 only1;
merge hold.editces1 (in=a) rangroup (in=b);
by ui;
if a & b then do;
      source_rg=1;
      output editces1;
end;
else if a then do;
      source_rg=2;
      output only1;
end;
run;

data editces1;
set editces1 only1;
rename source=source_cw;
run;

proc sort data=editces1;
by size;
run;

proc univariate data=editces1;
by size;
var ave_ae;
run;

proc freq data=editces1;
tables source_cw*source_rg;
run;

proc sort data=editces1 out=hold.editces2;
by naics_cw st size ces_id;
run;

proc contents data=hold.editces2;
run;
```

7. Analysis file creation

```
*options mprint;
Libname hold "c:\CES Data";

*Program Name: x:Research Project/Paper Programs/Analysis File1;
*Creates the analysis data file for Employment modeling and variances;

proc sort data=hold.ces_lopp out=ces_lopp;
by ces_id;
run;

proc sort data=hold.editces2 out=editces2;
by ces_id;
run;

data ces;
merge ces_lopp editces2;
by ces_id;
run;

proc sort data=ces;
by ui;
run;

data rangroup;
set hold.rangroup (drop=grandfl subsplwt);
run;

proc sort data=rangroup nodupkey;
by ui;
run;

data editces1;
merge ces (in=a) rangroup (in=b);
by ui;
if a & b then output editces1;
else if a then output editces1;
run;

%macro keep1(mon);
data
%do a=3 %to &mon;
        %do b=&a-1 %to &a-1;
                %do c=&a-2 %to &a-2;
```

```
                        ces&a (keep=ind size selwt lopp first_mo h1-h80
                                LR&a NR&a atyp&a ae&a LR&b NR&b atyp&b ae&b
LR&c NR&c ae&c)
                %end;
        %end;
%end;
;
%mend;

%macro keep2(mon);
set editces1 (keep=naics_cw size selwt first_mo lopp h1-h80
%do a=1 %to &mon;
        LR&a NR&a atyp&a ae&a
%end;
);
%mend;

%macro pull(mon);
%do a=3 %to &mon;
        %do b=&a-1 %to &a-1;
                %do c=&a-2 %to &a-2;
                        data ces&a (drop=first_mo);
                        set ces&a;
                                if first_mo le &b;
                                month=&a;
                                rename LR&a=LR_0;
                                rename NR&a=NR_0;
                                rename ae&a=y_0;
                                rename atyp&a=atyp_0;
                                rename LR&b=LR_1;
                                rename NR&b=NR_1;
                                rename ae&b=y_1;
                                rename atyp&b=atyp_1;
                                rename LR&c=LR_2;
                                rename NR&c=NR_2;
                                rename ae&c=y_2;
                %end;
        %end;
%end;
%mend;


%macro combine(mon);
set
%do a=3 %to &mon;
        ces&a
```

```
%end;
;
%mend;

%keep1(36);
%keep2(36);
if naics_cw="cons" then ind=1;
else if naics_cw="mfg " then ind=2;
else if naics_cw="mini" then ind=3;
else if naics_cw="whol" then ind=4;
run;

%pull(36);
run;


data analysis1;
%combine(36);
run;

proc contents data=analysis1;
run;

data hold.analysis1;
set analysis1;
run;

*Remember to add n=1 when doing Resp Status Modeling;
*Remember to create dummy variables when doing modeling;
*Remember to recode LR and NR when doing summary counts;

Proc means data=est2;
class ind month emp1;
var err_PR err_est err_model err_PRpct err_estpct err_modelpct;
title "summary of absolute errors for late reporters - LR with 1+ prior emp, unknown
prior emp change";
run;
```

8. Summarize LDB information to obtain benchmark counts

```
Filename ldb01 "c:\CES Data\ldb12863.dat";
Filename ldb02 "c:\CES Data\ldb12867.dat";
libname hold "c:\CES Data";

*Program Name: x:/Research Project/Paper Programs/LDB - Links Analysis
*This program reads the LDB extract files,
*assigns a size class
*and outputs a file with benchmark data for 2001, 2002;

*read LDB data for 2001;

data ldb01 (drop=naics_ldb);
infile ldb01 missover;
input @1 ldb 9.
            @10 state 2.
    @12 ui 10.
    @22 run 5.
            @32 naics_ldb 6.
            @38 emp01 6.
    @;
if naics_ldb ge 440000 then delete;
else if naics_ldb ge 420000 then naics="whol";
else if naics_ldb ge 310000 then naics="mfg ";
else if naics_ldb ge 230000 then naics="cons";
else if naics_ldb ge 220000 then delete;
else if naics_ldb ge 210000 then naics="mini";
else if naics_ldb ge 114000 then delete;
else if naics_ldb ge 113300 then naics="mini";
else delete;
      if emp01 le 10 then size=1;
      else if emp01 le 20 then size=2;
      else if emp01 le 50 then size=3;
      else if emp01 le 100 then size=4;
      else if emp01 le 150 then size=5;
      else if emp01 le 500 then size=6;
      else if emp01 le 1000 then size=7;
      else size=8;
run;

proc sort data=ldb01;
by state ui run;
run;
```

```
*read LDB data for 2002;

data ldb02 (drop=naics_ldb);
infile ldb02 missover;
input @1 ldb 9.
                    @10 state 2.
      @12 ui 10.
      @22 run 5.
                    @32 naics_ldb 6.
                    @38 emp02 6.
      @;
run;

proc sort data=ldb02;
by state ui run;
run;

data ldb only01 only02;
merge ldb01 (in=a) ldb02 (in=b);
by state ui run;
if a & b then output ldb;
else if a then output only01;
else if b then output only02;
run;

proc sort data=ldb;
by naics size;
run;

proc summary data=ldb noprint;
          class naics size;
          var emp01 emp02;
          output out=tot_ldb sum=emp01 emp02;
run;

data tot_ldb (drop=_type_ _freq_);
set tot_ldb;
if _type_ ge 2;
run;

proc sort data=tot_ldb out=hold.tot_ldb;
by naics size;
run;

proc print data=hold.tot_ldb;
run;
```

D. Coefficient Estimates for Full Logit Late Reporting Probability Model

Logit Model for Conditional Probability of Late Reporting
Coefficient Estimates

		Construction			Manufacturing			Mining			Wholesale Trade		
		2.5% Level	Estimate	97.5% Level	2.5% Level	Estimate	97.5% Level	2.5% Level	Estimate	97.5% Level	2.5% Level	Estimate	97.5% Level
Intercept		0.7380	0.8005	0.8582	0.9015	0.9818	1.0576	0.6934	0.8185	0.9591	0.9965	1.0735	1.1570
Number of Reporting Days	9	0.0855	0.1549	0.2240	0.0560	0.1303	0.2015	-0.0101	0.1720	0.3493	-0.1958	-0.1127	-0.0289
	10	-0.1169	-0.0723	-0.0216	-0.1545	-0.0981	-0.0382	-0.2427	-0.1187	-0.0070	-0.3575	-0.2961	-0.2359
	11	0.4122	0.4604	0.5177	0.0880	0.1400	0.1974	0.0401	0.1647	0.2877	-0.2195	-0.1486	-0.0820
	12*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	13	-0.1730	-0.1208	-0.0681	-0.0663	-0.0038	0.0622	-0.1986	-0.0676	0.0640	-0.4015	-0.3304	-0.2669
	14	-0.0188	0.0289	0.0796	-0.1540	-0.0984	-0.0490	-0.1332	-0.0047	0.1179	-0.2654	-0.1965	-0.1321
	15	0.0001	0.0728	0.1425	-0.1505	-0.0764	-0.0032	0.4106	0.6085	0.7982	1.0029	1.0982	1.1851
Length of Pay Period	Weekly	-0.1860	-0.1360	-0.0915	-0.2164	-0.1799	-0.1399	-0.0670	0.0084	0.0909	-0.1879	-0.1449	-0.1003
	Bi-Weekly*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Semi-Monthly	-0.4800	-0.4067	-0.3324	-0.2607	-0.2064	-0.1431	-0.2135	-0.0956	0.0237	-0.2134	-0.1604	-0.1111
	Monthly	0.2281	0.3126	0.3923	0.0796	0.1574	0.2350	0.1114	0.2467	0.3980	0.2753	0.3368	0.3952
Design Size Class	<10	-0.6761	-0.6359	-0.5910	-0.5116	-0.4226	-0.3376	-0.6693	-0.5312	-0.3656	-0.4658	-0.4126	-0.3400
	10-19	-0.4622	-0.4119	-0.3611	-0.4733	-0.3836	-0.2899	-0.4059	-0.2457	-0.0849	-0.3002	-0.2128	-0.1295
	20-49	-0.3156	-0.2693	-0.2230	-0.2723	-0.2016	-0.1326	-0.2114	-0.0789	0.0500	-0.1929	-0.1179	-0.0417
	50-99*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	100-249	0.0034	0.0474	0.0931	0.0984	0.1549	0.2082	0.0696	0.2064	0.3354	0.0837	0.1574	0.2257
	250-499	-0.0318	0.0305	0.0969	0.1192	0.1754	0.2329	-0.1786	-0.0399	0.1031	0.0236	0.1085	0.1811
	500-999	-0.1591	-0.0716	0.0241	0.0739	0.1446	0.2131	-0.2366	-0.0840	0.0674	0.0593	0.1518	0.2362
	1000+	-0.0046	0.0860	0.1794	0.2000	0.2577	0.3110	0.0833	0.2334	0.3814	0.2503	0.3242	0.3901
Reporting Pattern	$X^{PR}_{(t-1)si} = 1$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$X^{LR}_{(t-1)si} = 1$	0.5616	0.6019	0.6413	0.6053	0.6437	0.6844	0.7026	0.7881	0.8767	1.0370	1.0862	1.1340
	$X^{NR}_{(t-1)si} = 1$	-0.4717	-0.4219	-0.3678	-0.5487	-0.4967	-0.4450	-0.6608	-0.5019	-0.3637	-0.5315	-0.4611	-0.3869
	$X^{PR}_{(t-2)si} = 1$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$X^{LR}_{(t-2)si} = 1$	0.1321	0.1672	0.2052	0.1744	0.2100	0.2493	0.1371	0.2254	0.3183	0.2743	0.3154	0.3585
	$X^{NR}_{(t-2)si} = 1$	-0.4361	-0.3635	-0.3344	-0.8052	-0.7588	-0.7106	-0.5494	-0.4265	-0.3116	-0.7039	-0.6483	-0.5913
Reporting Gap		-1.4880	-1.4416	-1.3970	-1.4700	-1.4279	-1.3845	-1.6500	-1.5370	-1.4120	-1.7620	-1.6973	-1.6390

E. Notes Concerning Variance Estimation for Predicted Conditional Late Reporting Rates

Reporting Status Model

$$i \ni \left(\delta_i = 1, X_{ti}^{PR} = 0\right)$$

$$X_{tci}^{LR} \mid X_{tci}^{PR} = 0 \sim Bin\left(1, p_{LR|X_{tci}^{PR}=0,ci}\right)$$

$$\text{logit}\left(p_{LR|X_{tci}^{PR}=0,ci}\right) =$$

$$\alpha_{\text{c}} + \boldsymbol{\gamma}_{(\text{t-1})c}^{T}\mathbf{X}_{(t-1)ci} + \boldsymbol{\gamma}_{(\text{t-2})c}^{T}\mathbf{X}_{(t-2)ci} + \gamma_{\text{Gc}}\ln\left(G_{(t-1)ci} + 1\right) + \boldsymbol{\gamma}_{\text{Sc}}^{T}\mathbf{S}_{ci} + \boldsymbol{\gamma}_{\text{Lc}}^{T}\mathbf{L}_{ci} + \boldsymbol{\gamma}_{\text{Dc}}^{T}\mathbf{D}_{t}$$

(for shorthand purposes, subscripts ci shortened to i)

$$X_{ti}^{LR} \mid X_{ti}^{PR} = 0 \sim Bin\left(1, p_{LR|X_{ti}^{PR}=0,i}\right)$$

$$\text{logit}\left(p_{LR|X_{ti}^{PR}=0,i}\right) =$$

$$\alpha + \boldsymbol{\gamma}_{(\text{t-1})}^{T}\mathbf{X}_{(t-1)i} + \boldsymbol{\gamma}_{(\text{t-2})}^{T}\mathbf{X}_{(t-2)i} + \gamma_{\text{G}}\ln\left(G_{(t-1)i} + 1\right) + \boldsymbol{\gamma}_{\text{S}}^{T}\mathbf{S}_{i} + \boldsymbol{\gamma}_{\text{L}}^{T}\mathbf{L}_{i} + \boldsymbol{\gamma}_{\text{D}}^{T}\mathbf{D}_{t}$$

$$= \boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}$$

Posterior mean of X_{ti}^{LR} for $i \ni \left(X_{ti}^{PR} = 0\right)$

$$E\left(X_{ti}^{LR} \mid \mathbf{X}_{sR}\right) = E\left[E\left(X_{ti}^{LR} \mid \mathbf{X}_{sR}, p_{LR|X_{ti}^{PR}=0,i}\right) \mid \mathbf{X}_{sR}\right]$$

$$= E\left(p_{LR|X_{ti}^{PR}=0,i} \mid \mathbf{X}_{sR}\right)$$

$$E\left(X_{ti}^{LR} \mid X_{ti}^{PR} = 0, \mathbf{X}_{sR}\right) = p_{LR|X_{ti}^{PR}=0,i} = \frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1 + \exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}$$

where $\mathbf{X}_{sR}$ = available sample reporting information

Estimate of X_{ti}^{LR} for $i \ni \left(X_{ti}^{PR} = 0\right)$

An estimate for X_{ti}^{LR}, $\hat{X}_{ti}^{LR,B}$, is obtained by approximating $E\left(p_{LR|X_{ti}^{PR}=0,i} \mid \mathbf{X}_{sR}\right)$ through MCMC methods, and substituting this approximation, $\hat{p}^{B}_{LR|X_{ti}^{PR}=0,i}$, into $E\left(Y_{ti} \mid \mathbf{Y}_{sR}\right)$

$$\hat{X}_{ti}^{LR,B} = \hat{p}^{B}_{LR|X_{ti}^{PR}=0,i} = \frac{\exp\left(\left(\hat{\boldsymbol{\gamma}}^{B}\right)^{T} \boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\left(\hat{\boldsymbol{\gamma}}^{B}\right)^{T} \boldsymbol{\Psi}_{ti}\right)}$$

Posterior variance of X_{ti}^{LR} for $i \ni \left(X_{ti}^{PR}=0\right)$

$$V\left(\hat{X}_{ti}^{LR} \mid X_{ti}^{PR}=0, \mathbf{X}_{sR}\right) = E\left\{V\left(\hat{X}_{ti}^{LR} \mid X_{ti}^{PR}=0, p_{LR|X_{ti}^{PR}=0,i}, \mathbf{X}_{sR}\right) \mid \mathbf{X}_{sR}\right\}$$

$$+V\left\{E\left(\hat{X}_{ti}^{LR} \mid X_{ti}^{PR}=0, \mathbf{X}_{sR}, p_{LR|X_{ti}^{PR}=0,i}\right) \mid \mathbf{X}_{sR}\right\}$$

$$= E\left\{p_{LR|X_{ti}^{PR}=0,i}\left(1-p_{LR|X_{ti}^{PR}=0,i}\right) \mid \mathbf{X}_{sR}\right\} + V\left\{p_{LR|X_{ti}^{PR}=0,i} \mid \mathbf{X}_{sR}\right\}$$

$$+V\left\{p_{LR|X_{ti}^{PR}=0,i} \mid \mathbf{X}_{sR}\right\}$$

$$= p_{LR|X_{ti}^{PR}=0,i}\left(1-p_{LR|X_{ti}^{PR}=0,i}\right) + V\left(\frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)} \mid \mathbf{X}_{sR}\right)$$

use Taylor series expansion to solve

$$\frac{\partial}{\partial \alpha}\left(\frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}\right) = \frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)} - \left(\frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}\right)^{2}$$

$$= \frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}\left(1-\frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}\right)$$

$$\frac{\partial}{\partial \gamma_{\mathrm{G}}}\left(\frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}\right) = \frac{\ln\left(G_{(t-1)i}+1\right)\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)} - \ln\left(G_{(t-1)i}+1\right)\left(\frac{\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^{T}\boldsymbol{\Psi}_{ti}\right)}\right)^{2}$$

$$= \frac{\ln\left(G_{(t-1)i}+1\right)\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}\left(1-\frac{\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}\right)$$

$$\frac{\partial}{\partial\gamma_F}\left(\frac{\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}\right)=\frac{F\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}-F\left(\frac{\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}\right)^2$$

$$=\frac{F\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}\left(1-\frac{\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}\right)$$

Where F represents any of the remaining parameters

$$V\left(\frac{\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}\right)=\left[p_{LR|X_{ti}^{PR}=0,i}\left(1-p_{LR|X_{ti}^{PR}=0,i}\right)\right]^2 V\left(\alpha\,|\,\mathbf{X}_{sR}\right)$$

$$+\left[p_{LR|X_{ti}^{PR}=0,i}\left(1-p_{LR|X_{ti}^{PR}=0,i}\right)\right]^2\left[\ln\left(G_{(t-1)i}+1\right)\right]V\left(\gamma_G\,|\,\mathbf{X}_{sR}\right)+\ldots$$

$$+\left[p_{LR|X_{ti}^{PR}=0,i}\left(1-p_{LR|X_{ti}^{PR}=0,i}\right)\right]^2\left[\ln\left(G_{(t-1)i}+1\right)\right]Cov\left(\alpha,\gamma_G\,|\,\mathbf{X}_{sR}\right)+\ldots$$

$$=\left(p_{LR|X_{ti}^{PR}=0,i}\left(1-p_{LR|X_{ti}^{PR}=0,i}\right)\boldsymbol{\Psi}_{ti}\right)^T\mathbf{V}\left(\boldsymbol{\gamma}\,|\,\mathbf{X}_{sR}\right)\left(p_{LR|X_{ti}^{PR}=0,i}\left(1-p_{LR|X_{ti}^{PR}=0,i}\right)\boldsymbol{\Psi}_{ti}\right)$$

Estimated Variance

An estimate of the variance of $\hat{X}_{ti}^{LR}$, $v^B\left(X_{ti}^{LR}\right)$, is obtained by approximating $V\left(\boldsymbol{\gamma}\,|\,\mathbf{X}_{sR}\right)$ and $p_{LR|X_{ti}^{PR}=0,i}$ through MCMC methods, and substituting the approximations, $\mathbf{V}^B\left(\boldsymbol{\gamma}\right)$ and $p_{LR|X_{ti}^{PR}=0,i}$, respectively, into $V\left(\frac{\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}{1+\exp\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_{ti}\right)}\right)$

$$V\left(\hat{X}_{ti}^{LR}\,|\,X_{ti}^{PR}=0,\mathbf{X}_{(t-1)i},\mathbf{X}_{(t-2)i},\mathbf{Z},\mathbf{X}_{sR}\right)$$

$$=\left(\hat{p}_{LR|X_{ti}^{PR}=0,i}\left(1-\hat{p}_{LR|X_{ti}^{PR}=0,i}\right)\boldsymbol{\Psi}_{ti}\right)^T V\left(\hat{\boldsymbol{\gamma}}\right)\left(\hat{p}_{LR|X_{ti}^{PR}=0,i}\left(1-\hat{p}_{LR|X_{ti}^{PR}=0,i}\right)\boldsymbol{\Psi}_{ti}\right)$$

F. Selected Code for Reporting Status Model Implementation

1. Model specification for WinBUGS

```
model {
  for (i in 1:N){
    LR[i] ~ dbin (p[i],n[i])
    logit(p[i]) <- a + inprod(gD[],days[i,]) + gL*lopp[i] + inprod(gS[],size[i,]) +
        gLR1*LR1[i] + gLR2*LR2[i] + gNR1*NR1[i] + gNR2*NR2[i] + gG*gap[i]

  }

  a ~ dunif (-5, 5)
  gLR1 ~ dunif (-5, 5)
  gLR2 ~ dunif (-5, 5)
  gNR1 ~ dunif (-5, 5)
  gNR2 ~ dunif (-5, 5)
  gD[1] ~ dunif (-2, 2)
  gD[2] ~ dunif (-2, 2)
  gL ~ dunif (-2, 2)
  gS[1] ~ dunif (-2, 2)
  gS[2] ~ dunif (-2, 2)
  gS[3] ~ dunif (-2, 2)
  gG ~ dunif (-5, 5)

}
```

2. R code used to read data and call WinBUGS

```
# response status model with collapsed factors - months 3 - k
############################
#Mining
############################
#months 3-15

test3 <- read.table ("LR1_mine3.txt", header=T)
test4 <- read.table ("LR1_mine4.txt", header=T)
test5 <- read.table ("LR1_mine5.txt", header=T)
test6 <- read.table ("LR1_mine6.txt", header=T)
test7 <- read.table ("LR1_mine7.txt", header=T)
test8 <- read.table ("LR1_mine8.txt", header=T)
test9 <- read.table ("LR1_mine9.txt", header=T)
test10 <- read.table ("LR1_mine10.txt", header=T)
test11 <- read.table ("LR1_mine11.txt", header=T)
test12 <- read.table ("LR1_mine12.txt", header=T)
test13 <- read.table ("LR1_mine13.txt", header=T)
test14 <- read.table ("LR1_mine14.txt", header=T)
test15 <- read.table ("LR1_mine15.txt", header=T)
test <- rbind(test3, test4, test5, test6, test7, test8, test9, test10, test11, test12, test13,
test14, test15)
N <- nrow(test)
n <- test$n
LR <- test$LR
LR1 <- test$LR1
LR2 <- test$LR2
NR1 <- test$NR1
NR2 <- test$NR2
days.1 <- test$days9 + test$days11
days.2 <- test$days15
days <- cbind(days.1, days.2)
size.1 <- test$size1
size.2 <- test$size2
size.3 <- test$size5 + test$size8
size <- cbind(size.1, size.2, size.3)
lopp <- test$lopp4
gap <- log(test$gap)

data <- list("N", "n", "LR", "LR1", "LR2", "NR1", "NR2", "days", "lopp", "size",
"gap")
inits1 <- list(a=1.1, gD=c(0.1,0.1), gL=0.1, gS=c(0.1,0.1,0.1), gG=0.1, gLR.1=0.1,
gLR.2=0.1, gNR.1=0.1, gNR.2=0.1)
```

```
    inits2 <- list(a=0.9, gD=c(-0.1,-0.1), gL=-0.1, gS=c(-0.1,-0.1,-0.1), gG=-0.1,
gLR.1=-0.1, gLR.2=-0.1, gNR.1=-0.1, gNR.2=-0.1)
    inits <- list(inits1, inits2)
    parameters <- c("a", "gD", "gL", "gS", "gLR1", "gLR2", "gNR1", "gNR2", "gG")
    test.sim <- bugs (data, inits, parameters, "Resp Status Model Initial Extended-
Mine.txt", n.chains=2, n.iter=500, digits=4)

    attach.all(test.sim)
    test.sim$summary

    Mean<-test.sim$summary[1:13,1]
    Stdev<-test.sim$summary[1:13,2]
    Per2.5<-test.sim$summary[1:13,3]
    Per97.5<-test.sim$summary[1:13,7]
    Rhat<-test.sim$summary[1:13,8]
    n.eff<-test.sim$summary[1:13,9]
    DIC.16 <- DIC
    pD.16 <- pD

    results16<-
data.frame(M16=Mean,SD16=Stdev,LPer16=Per2.5,UPer16=Per97.5,R16=Rhat,n16
=n.eff)
    params16<-data.frame(M16=Mean)
```

3. Summary of results of model

```
*options mprint;
Libname source "c:\unzipped";

data cons;
set source.cons_lr1;
if month ge 16;
if month le 27;
run;

proc sort data=cons;
by month;
run;

data parameters;
infile 'I:\Bayes\CES\Paper\parconsrev3.csv' delimiter=',';
input month a gD1 gD2 gD3 gL1 gL2 gL3 gS1 gS2 gS3 gLR1 gLR2 gNR1 gNR2
gG;
run;

proc sort data=parameters;
by month;
run;

data pred;
merge cons parameters;
by month;
logit = a + gG*log(gap)
        + gS1*size1 + gS2*size2 + gS3*size3
        + gL1*lopp4 + gL2*lopp2 + gL3*lopp3
        + gD1*(days10 + days13) + gD2*days9 + gD3*days11
        + gLR1*LR_1 + gNR1*NR_1 + gLR2*LR_2 + gNR2*NR_2;
Pred_Prior=exp(logit)/(1 + exp(logit));
run;

*proc print data=pred;
*run;

proc summary data=pred noprint;
class month LR_1 NR_1 LR_2 NR_2;
var n LR Pred_Prior;
output out=Results sum = Total LR Pred_Prior;
run;

*proc print data=results;
*run;
```

```
data Results (drop=_type_ _freq_);
set Results;
if _type_=16 or _type_=31;
Actual_LR=LR/Total;
Pred_LR_Prior=Pred_Prior/Total;
Err_Prior=Pred_LR_Prior-Actual_LR;
run;

proc sort data=Results;
by LR_1 NR_1 LR_2 NR_2 month;
run;

data parameters16 (drop=month);
set parameters;
if month=16;
n=1;
run;

proc sort data=parameters16;
by n;
run;

proc sort data=cons;
by n;
run;

data pred16;
merge parameters16 cons;
by n;
logit = a + gG*log(gap)
        + gS1*size1 + gS2*size2 + gS3*size3
        + gL1*lopp4 + gL2*lopp2 + gL3*lopp3
        + gD1*(days10 + days13) + gD2*days9 + gD3*days11
        + gLR1*LR_1 + gNR1*NR_1 + gLR2*LR_2 + gNR2*NR_2;
Pred_Prior=exp(logit)/(1 + exp(logit));
run;

*proc print data=pred;
*run;

proc summary data=pred16 noprint;
class month LR_1 NR_1 LR_2 NR_2;
var n LR Pred_Prior;
output out=Results16 sum = Total LR Pred_Prior;
run;
```

```
*proc print data=results;
*run;

data Results16 (drop=_type_ _freq_ LR Total Pred_Prior);
set Results16;
if _type_=16 or _type_=31;
Actual_LR=LR/Total;
Pred_LR_Prior16=Pred_Prior/Total;
Err_Prior16=Pred_LR_Prior16-Actual_LR;
run;

proc sort data=Results16;
by LR_1 NR_1 LR_2 NR_2 month;
run;

data Results;
merge Results Results16;
by LR_1 NR_1 LR_2 NR_2 month;
run;

proc print data=Results;
var month LR_1 NR_1 LR_2 NR_2 Total LR Actual_LR Pred_LR_Prior Err_Prior
Pred_LR_Prior16 Err_Prior16;
title "LR Prediction Results for Construction for Months 16-24";
run;

PROC EXPORT DATA= WORK.RESULTS
            OUTFILE= "I:\Bayes\CES\Paper Results\Cons Pred LR.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;
```

G. Bayes' Estimation of Fixed Effects

Parameters for the employment growth models were estimated using SAS v.8.2 for the current method and Model 2, and SAS v.8.2 and WinBUGS v.1.4 called from a program written in R v.1.8.1 for the hierarchical fixed effect approach (Model 2). The WinBUGS model specification is provided in Appendix H.1. The R code used for calling WinBUGS is provided in Appendix H.2. Missing employment was imputed under Models 1 and 2 for sample units reporting in month $t-1$ that had not reported data for month t in time for preliminary estimation.

The sample for three of the industries (Construction, Manufacturing, and Wholesale Trade), was on the order of 10 to 20 times as large as that for the remaining industry (Mining). As a result, the WinBUGS program for these industries had a run time over 24 hours for one month within an industry (versus approximately two hours for Mining). In order to provide a more efficient run time, these industries were subsampled at a 10% rate (for Construction and Wholesale Trade) or a 5% rate (for Manufacturing). Even with these reductions in sample size, each model ran on the order of 1-2 hours. Given the model was run for 12 months for each of 4 industries, the computing time required to obtain all the necessary parameter estimates was on the order of 3-4 days (not including the inevitable glitches involved in testing the program code).

The model was run using two chains, with 200 iterations and a burn-in period of 100 iterations. Initial values for each parameter were set at 0.1 above the mean for the distribution for chain one and 0.1 below the mean for the distribution for chain two. Averages for the potential scale reduction factors for the model across the 12 months are provided in Table 46. As can be seen, there were four parameters that failed to meet the guideline convergence criteria for at least one month. However, further examination showed occurrences were at most two for any parameter, so the model was not run using additional iterations.

Table 46-PSRF Values for Employment Growth Model 2

Maximum Potential Scale Reduction Factors for Model II
March 2001 - April 2002

	Construction		Manufacturing		Mining		Wholesale Trade	
	Prior Month Size Class							
	Small	Large	Small	Large	Small	Large	Small	Large
$\rho_{(t-6)c}$	1.05	1.19	1.02	1.16	1.05	1.19	1.02	1.02
$\rho_{(t-5)c}$	1.03	1.11	1.02	1.39	1.03	1.04	1.02	1.11
$\rho_{(t-4)c}$	1.02	1.05	1.03	1.07	1.04	1.07	1.02	1.07
$\rho_{(t-3)c}$	1.07	1.15	1.04	1.18	1.09	1.14	1.05	1.08
$\rho_{(t-2)c}$	1.02	1.16	1.05	1.20	1.03	1.08	1.05	1.03
$\rho_{(t-1)c}$	1.03	1.12	1.03	1.10	1.03	1.09	1.02	1.07
$\lambda_c^{(low)}$	1.07	1.18	1.05	1.23	1.10	1.26	1.04	1.14
$\lambda_c^{(high)}$	1.11	1.19	1.02	1.35	1.02	1.11	1.02	1.14
$\lambda_c^{(unk)}$	1.06	1.15	1.04	1.17	1.04	1.08	1.02	1.05
σ_y	1.10	1.07	1.04	1.05	1.04	1.04	1.02	1.07

Several illustrations from the graphical results available from the R software used to call WinBUGS are provided in Figure 35-Figure 37. The parameter, “rho[k]” corresponds to the proportionality factor for month $t-(7-k)$ (i.e., months were sequentially ordered in the WinBUGS specification from 1 to 6, with 1 representing the oldest month, $t-6$, and 6 representing the most recent month, $t-1$), “pC[k]” corresponds to the $\lambda's$ $\left(pC[1]\Rightarrow\lambda^{(low)},\ pC[2]\Rightarrow\lambda^{(high)},\ pC[3]\Rightarrow\lambda^{(unk)}\right)$. Looking at the graphs, the greater variability associated with the estimate for $\lambda^{(unk)}$ can be seen. Refer to Chapter IV for an explanation of the structure of the graphs.

Figure 35-Model 2 Results for March 2002: Manufacturing, Large Employment

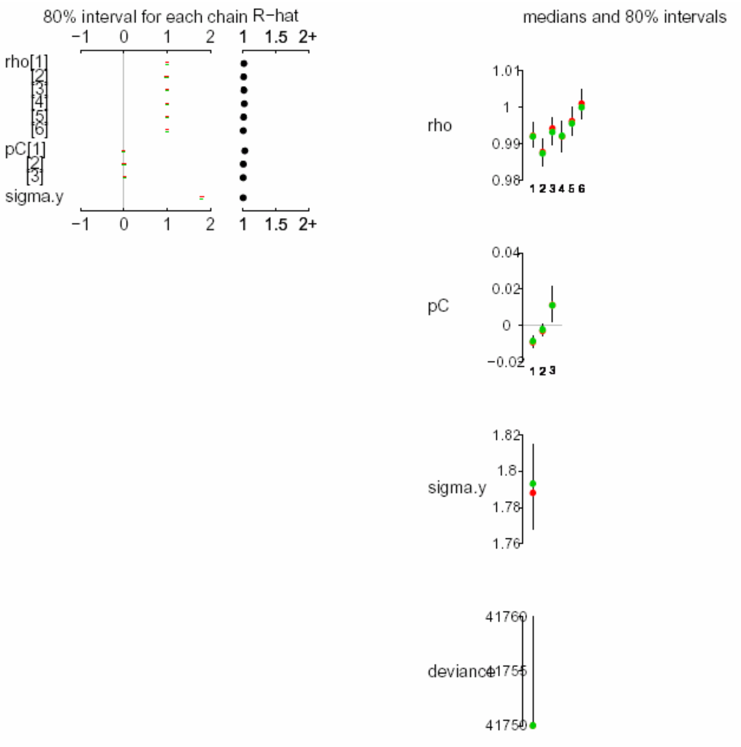


Figure 36-Model 2 Results for March 2002: Mining, Large Employment

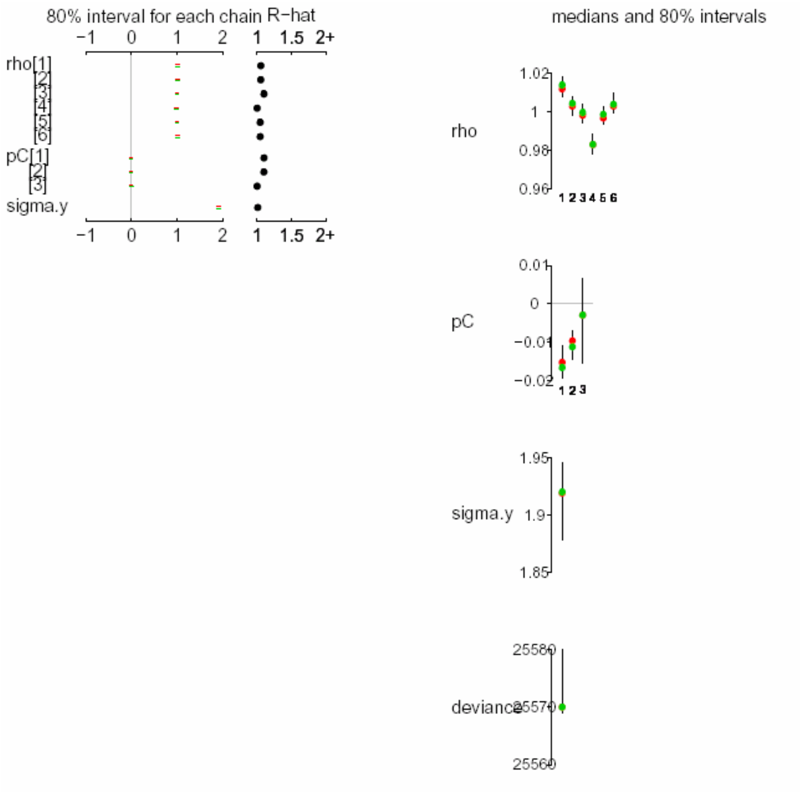
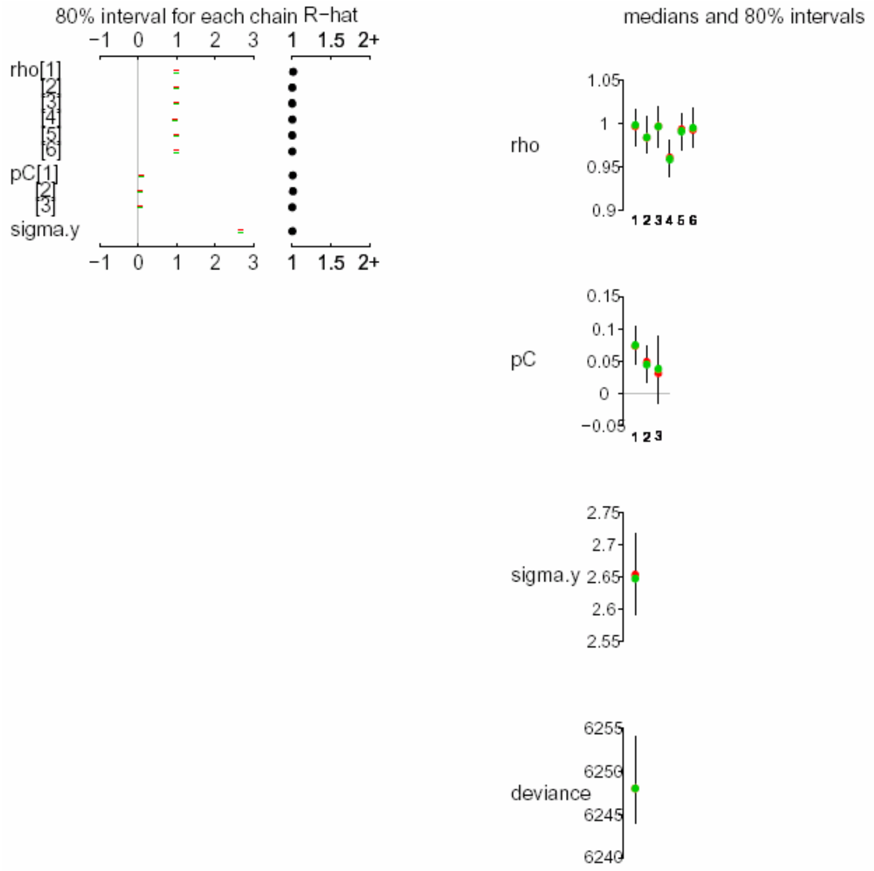


Figure 37-Model 2 Results for March 2002: Mining, Small Employment



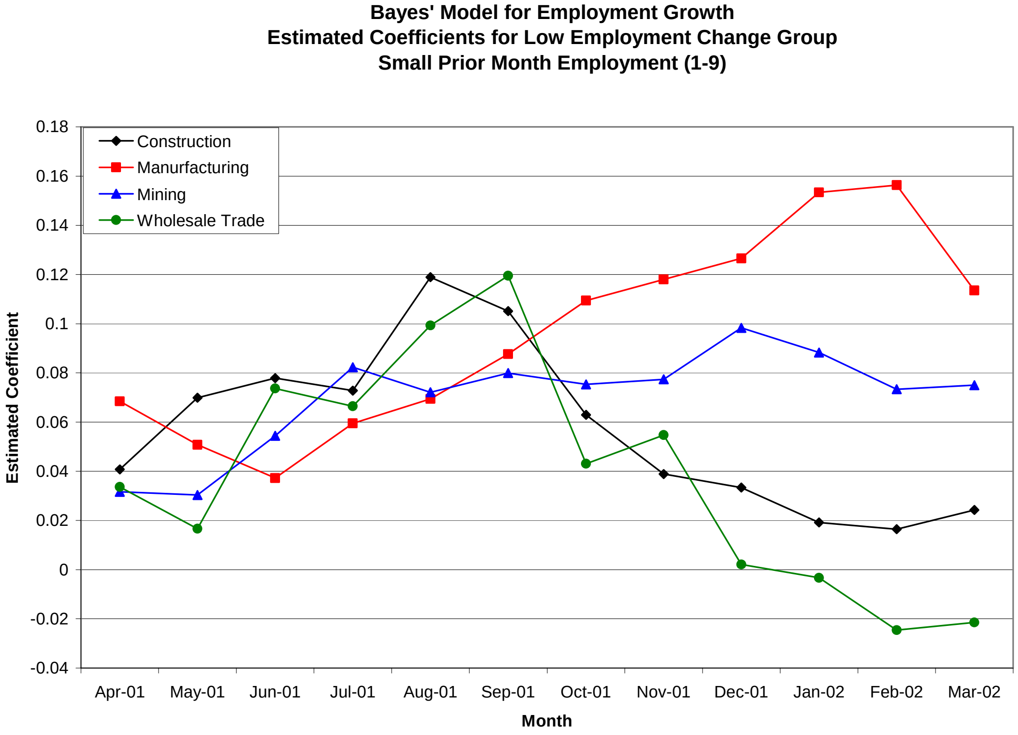
Values for the estimated parameters are relatively unstable for the small prior employment group, but fairly stable across time for the large prior employment group, as indicated in Table 47 and Figure 38. The standard deviations were used rather than a relative standard deviation, as the coefficients are roughly equivalent. It should also be noted that many of the coefficients are not significantly different from zero, which is somewhat to be expected as, based on the review of link relatives by characteristic discussed in Chapter V, deviations from the industry level for a group are expected to be relatively small.

Table 47-Distribution of Coefficient Estimates for Model 2

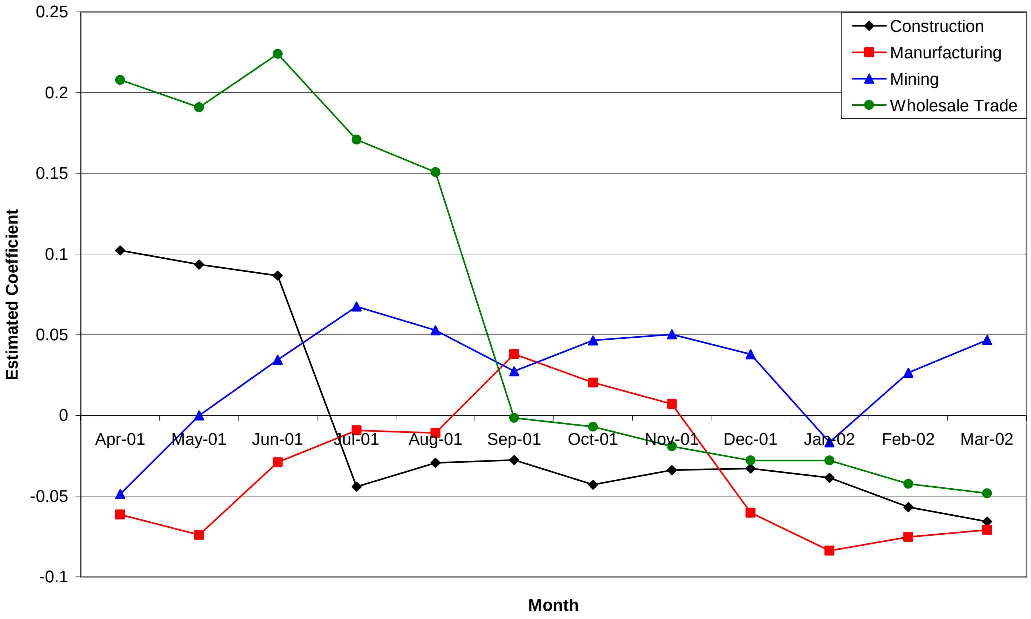
Distribution of Coefficient Estimates
April 2001 - March 2002

Prior Employment Size	Employment Growth Group	Construction		Manufacturing		Mining		Wholesale Trade	
		Mean	st dev	Mean	st dev	Mean	st dev	Mean	st dev
Small (1-9)	Low	0.06	0.03	0.10	0.04	0.07	0.02	0.04	0.05
	High	-0.01	0.06	-0.03	0.04	0.03	0.03	0.06	0.11
	Unknown	-0.02	0.01	-0.01	0.01	0.00	0.04	-0.01	0.03
Large (10+)	Low	0.00	0.02	0.00	0.00	0.01	0.03	0.00	0.00
	High	0.01	0.01	0.01	0.01	0.00	0.02	0.00	0.01
	Unknown	0.01	0.02	0.02	0.01	0.01	0.05	0.00	0.01

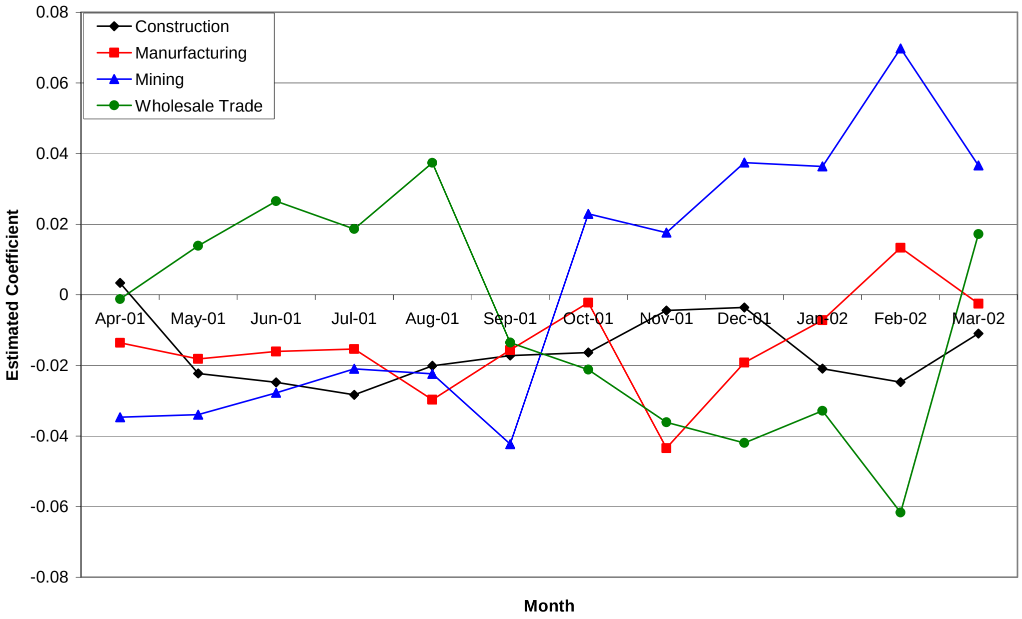
Figure 38-Coefficients for Model 2



Bayes' Model for Employment Growth
Estimated Coefficients for High Employment Change Group
Small Prior Month Employment (1-9)



Bayes' Model for Employment Growth
Estimated Coefficients for Unknown Employment Change Group
Small Prior Month Employment (1-9)



H. Selected Code for Employment Growth Model Implementation

1. Model specification for WinBUGS

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y[i])
    y.hat[i] <- (rho[month[i]] + inprod(pC[],change[i,]))*x[i]
    tau.y[i] <- pow(sigma.y, -2)*w[i]/z[i]
    z[i] <- max(1, x[i])
  }

  for (j in 1:n.month){
    rho[j] ~ dunif (.2, 1.8)
  }

  sigma.y ~ dunif (0, 1000)
  pC[1] ~ dunif (-2, 2)
  pC[2] ~ dunif (-2, 2)
  pC[3] ~ dunif (-2, 2)

}
```

2. R code used to read data and call WinBUGS

```
#########################################
# ae model
# assumes proportional relationship within size (small - <10, large, 10+)
# small
# months 10-15

test10 <- read.table ("smmine2_10.txt", header=T)
test11 <- read.table ("smmine2_11.txt", header=T)
test12 <- read.table ("smmine2_12.txt", header=T)
test13 <- read.table ("smmine2_13.txt", header=T)
test14 <- read.table ("smmine2_14.txt", header=T)
test15 <- read.table ("smmine2_15.txt", header=T)
test <- rbind(test10, test11, test12, test13, test14, test15)
n <- nrow(test)
n.month <- max(test$month)-9

x <- test$x
y <- test$y
w <- test$selwt
month <- test$month-9
change1 <- test$change1
change2 <- test$change2
change3 <- test$change3
change <- cbind(change1,change2,change3)

data <- list("n", "n.month", "x", "y", "w", "month", "change")
inits1 <- list(rho=c(1.01,1.01,1.01,1.01,1.01,1.01), pC=c(0.01,0.01,0.01),
sigma.y=0.1)
inits2 <- list(rho=c(0.99,0.99,0.99,0.99,0.99,0.99), pC=c(-0.01,-0.01,-0.01),
sigma.y=0.1)
inits <- list(inits1, inits2)
parameters <- c("rho", "pC", "sigma.y")
test.sim <- bugs (data, inits, parameters, "Large AE Model.txt", n.chains=2,
n.iter=200, digits=4)

attach.all(test.sim)
test.sim$summary
Mean<-test.sim$summary[1:11,1]
Stdev<-test.sim$summary[1:11,2]
LPer<-test.sim$summary[1:11,3]
UPer<-test.sim$summary[1:11,7]
Rhat<-test.sim$summary[1:11,8]
n.eff<-test.sim$summary[1:11,9]
```

```
res16s<-data.frame(M16s=Mean, Sd16s=Stdev, sPer16s=LPer, UPer16s=UPer,
R16s=Rhat, n16s=n.eff)
par16s<-data.frame(M16s=Mean)
DIC.16s<-DIC
pD.16s<-pD
temp<-data.frame(M=Mean, Sd=Stdev, sPer=LPer, UPer=UPer, R=Rhat, n=n.eff,
DIC=DIC, pD=pD)

write.table(temp, file= "Temp Mine2 Small 16.csv", sep= "," , col.names=NA)

########################################
# ae model
# assumes proportional relationship within size (small - <10, large, 10+)
# large
# months 10-15

test10 <- read.table ("lgmine2_10.txt", header=T)
test11 <- read.table ("lgmine2_11.txt", header=T)
test12 <- read.table ("lgmine2_12.txt", header=T)
test13 <- read.table ("lgmine2_13.txt", header=T)
test14 <- read.table ("lgmine2_14.txt", header=T)
test15 <- read.table ("lgmine2_15.txt", header=T)
test <- rbind(test10, test11, test12, test13, test14, test15)
n <- nrow(test)
n.month <- max(test$month)-9

x <- test$x
y <- test$y
w <- test$selwt
month <- test$month-9
change1 <- test$change1
change2 <- test$change2
change3 <- test$change3
change <- cbind(change1,change2,change3)

data <- list("n", "n.month", "x", "y", "w", "month", "change")
inits1 <- list(rho=c(1.01,1.01,1.01,1.01,1.01,1.01), pC=c(0.01,0.01,0.01),
sigma.y=0.1)
inits2 <- list(rho=c(0.99,0.99,0.99,0.99,0.99,0.99), pC=c(-0.01,-0.01,-0.01),
sigma.y=0.1)
inits <- list(inits1, inits2)
parameters <- c("rho", "pC", "sigma.y")
test.sim <- bugs (data, inits, parameters, "Large AE Model.txt", n.chains=2,
n.iter=200, digits=4)

attach.all(test.sim)
```

```
test.sim$summary
Mean<-test.sim$summary[1:11,1]
Stdev<-test.sim$summary[1:11,2]
LPer<-test.sim$summary[1:11,3]
UPer<-test.sim$summary[1:11,7]
Rhat<-test.sim$summary[1:11,8]
n.eff<-test.sim$summary[1:11,9]
res16l<-data.frame(M16l=Mean, Sd16l=Stdev, LPer16l=LPer, UPer16l=UPer,
R16l=Rhat, n16l=n.eff)
par16l<-data.frame(M16l=Mean)
DIC.16l<-DIC
pD.16l<-pD
temp<-data.frame(M=Mean, Sd=Stdev, sPer=LPer, UPer=UPer, R=Rhat, n=n.eff,
DIC=DIC, pD=pD)

write.table(temp, file= "Temp Mine2 Large 16.csv", sep= "," , col.names=NA)
```

3. Create imputed data and derive link relatives

```
*options mprint;
Libname hold "c:\CES Data";

*Program Name: x:Research Project/Paper Programs/Revision/Final;
*AE Estimation variance-4size;
*Calculates full and half-sample estimates;

%macro createvar(var);
p0=prior;
c0=curr;
p0p=prior;
c0p=curr;
p0f=prior;
c0f=curr;
p1=prior;
c1=curr;
p1f=prior;
c1f=curr;
%mend;

*Calculate full sample estimates;

data all (keep=ind month size n ch selwt curr prior emp1 group2 LR_0
NR_0 NR_1 NR_2 R);
set hold.analysis1r (keep=LR_0 NR_0 NR_1 NR_2 atyp_0 atyp_1
      ind month y_0 y_1 y_2 selwt size);
if NR_1 = 1 then delete;
if atyp_0 ge 1 then delete;
if y_1 = . then delete;
n=1;
if atyp_1 ge 1 then do;
      NR_2=1;
      y_2=.;
end;
if selwt = . then selwt=1;
if y_1 le 9 then emp1=1;
      else if y_1 le 19 then emp1=2;
      else if y_1 le 49 then emp1=3;
      else emp1=4;
if y_2 = . then group2=0;
      else if y_2 le 9 then group2=1;
      else if y_2 le 19 then group2=2;
      else if y_2 le 49 then group2=3;
      else if y_2 le 99 then group2=4;
      else if y_2 le 249 then group2=5;
      else if y_2 le 499 then group2=6;
      else group2=7;
if atyp_1 ge 1 then ch=.;
if NR_2=0 then ch=y_1-y_2;
rename y_1=prior;
rename y_0=curr;
if LR_0=1 then R=2;
      else if NR_0=1 then R=3;
      else R=1;
run;

*Group into quintiles by ch-within ind, month month t-2 emp group;
*yield low=0, med=1, hi=2, unk=.;
proc sort data=all;
by ind month group2;
run;

proc rank data=all out=all groups=3;
by ind month group2;
var ch;
ranks ch_r;
run;

*rename change to make low=1, med=2, hi=3, unk=4;
*use actual change for month t-1 emp <10, relative change for month
t-1 emp 10+
*create dummy variables for use in model estimation;
data all (drop=ch);
set all;
if emp1 le 3 then rch=ch_r;
      else rch=.;
if rch ge 0 then do;
      if rch = 0 then change=1;
      else if rch =1 then change=2;
      else if rch = 2 then change=3;
end;
else change=4;
run;

*Create subsets for use in estimating LRs, imputation;
*conf1: LR in month t, R in month t-1 and t-2, month t-1 emp>0;
*for1: not LR in month t, R in month t-1 and t-2, month t-1 emp>0;
*q data sets should be empty;
data conf1 conq1 only0
      for1 forq1
      Q1;
set all;
if emp1=1 then do;
      if change=1 then cell=1;
      else if change=3 then cell=2;
      else cell=3;
end;
else if emp1=2 then do;
      if change=1 then cell=4;
      else if change=3 then cell=5;
      else cell=6;
end;
else if emp1=3 then do;
      if change=1 then cell=7;
      else if change=3 then cell=8;
      else cell=9;
end;
else cell=10;
if NR_0+LR_0=0 then do;
      if NR_1=0 then do;
```

```
            if prior ge 0 then output conf1;
            else output conq1;
      end;
      else output only0;
end;
else if NR_0+LR_0=1 then do;
      if prior ge 0 then output for1;
      else output forq1;
end;
else output Q1;
run;

*calculate current link relatives;
*preliminary;
*combine PRs that reported in month t-1;
data PR;
set conf1;
run;

proc sort data=PR;
by ind month;
run;

proc summary data=PR nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_PR sum=p c;
run;

data LR_PR (drop=p c);
set LR_PR (drop=_type_ _freq_);
lr_PR=c/p;
run;

proc sort data=LR_PR;
by ind month;
run;

*final;
*combine LRs that reported in month t-1 with PRs that reported in
month t-1;
data LR;
set for1;
if LR_0=1;
run;

data Rpt;
set PR LR;
run;

proc sort data=Rpt;
by ind month;
run;

proc summary data=Rpt nway;
by ind month;
```

```
var prior curr;
weight selwt;
output out=LR_Rpt sum=p c;
run;

data LR_Rpt (drop=p c);
set LR_Rpt (drop=_type_ _freq_);
lr_R=c/p;
run;

proc sort data=LR_Rpt;
by ind month;
run;

*First level estimation;
*create variables for use in comparions for constant reporters;
*0 refers to reported values;
*1 refers to current size x prior change imputed values;
*f refers to final values;

data conf1;
set conf1;
%createvar(1);
run;

*Carry out imputation;
***********;
*month t-1 Rpt;
*calcuate link relative for size x change;
proc sort data=conf1;
by ind month cell;
run;

proc summary data=conf1 nway;
by ind month cell;
var p0 c0;
weight selwt;
output out=LR_sc sum=p c;
run;

data LR_sc (drop=p c);
set LR_sc (drop=_type_ _freq_);
lr_sc=c/p;
run;

proc sort data=LR_sc;
by ind month cell;
run;

*merge with ind link relatives;
data LR_conf1;
merge LR_PR LR_Rpt LR_sc;
by ind month;
run;

*impute for missing values using emp x change, ind, model;
*missing month t, month t-1 emp>0, prior change available;
```

```
proc sort data= for1;
by ind month cell;
run;

data for1adj;
merge LR_conf1 for1 (in=a);
by ind month cell;
if a;
c1=prior*lr_sc;
p1=prior;
if LR_0=1 then do;
      p0p=prior;
      c0p=prior*lr_PR;
      p0f=prior;
      c0f=curr;
      c1f=curr;
      p1f=prior;
end;
else if NR_0=1 then do;
      c1f=c1;
      p1f=p1;
end;
run;

*Final estimation;
*use all available records;

data imp;
set conf1 for1adj;
run;

proc sort data=imp;
by ind month LR_0;
run;

proc summary data=imp nway;
by ind month;
var p0 c0 p0f c0f p1 c1 p1f c1f;
weight selwt;
output out=LR_imp_all sum=p0 c0 p0f c0f pI cI pIf cIf;
run;

data LR_imp_all (drop=p0 c0 p0f c0f pI cI pIf cIf);
set LR_imp_all (drop=_type_ _freq_);
lr0=c0/p0;
lr0f=c0f/p0f;
lr1=cI/pI;
lr1f=cIf/pIf;
run;

proc sort data=LR_imp_all;
by ind month;
run;
```

4. Derive balanced half sample estimates

```
*Calculate half-sample estimates;

%macro rg(rg);
%do z=1 %to &rg;
data all_neg (keep=ind month size n ch selwt curr prior emp1 group2
LR_0 NR_0 NR_1 NR_2)
      all_pos (keep=ind month size n ch selwt curr prior emp1 group2
LR_0 NR_0 NR_1 NR_2);
set hold.analysis1r (keep=LR_0 NR_0 NR_1 NR_2 atyp_0 atyp_1 h&z
      ind month y_0 y_1 y_2 selwt size);
if h&z=. then delete;
if NR_1 = 1 then delete;
if atyp_0 ge 1 then delete;
if y_1 = . then delete;
n=1;
if atyp_1 ge 1 then do;
      NR_2=1;
      y_2=.;
end;
if selwt = . then selwt=1;
selwt=1+0.5*h&z;
if y_1 le 9 then emp1=1;
      else if y_1 le 19 then emp1=2;
      else if y_1 le 49 then emp1=3;
      else emp1=4;
if y_2 = . then group2=0;
      else if y_2 le 9 then group2=1;
      else if y_2 le 19 then group2=2;
      else if y_2 le 49 then group2=3;
      else if y_2 le 99 then group2=4;
      else if y_2 le 249 then group2=5;
      else if y_2 le 499 then group2=6;
      else group2=7;
if atyp_1 ge 1 then ch=.;
if NR_2=0 then ch=y_1-y_2;
rename y_1=prior;
rename y_0=curr;
if h&z=-1 then output all_neg;
else if h&z=1 then output all_pos;
run;

****************;
*run for one half sample;
*Group into tertiles by ch-within ind, month month t-2 emp group;
*yield low=0, med=1, hi=2, unk=.;
proc sort data=all_neg;
by ind month group2;
run;

proc rank data=all_neg out=all_neg groups=3;
by ind month group2;
var ch;
ranks ch_r;
```

```
run;

data all_neg (drop=ch);
set all_neg;
if emp1 le 3 then rch=ch_r;
      else rch=.;
if rch ge 0 then do;
      if rch = 0 then change=1;
      else if rch =1 then change=2;
      else if rch = 2 then change=3;
end;
else change=4;
run;

data conf1 conq1 only0
      for1 forq1
      Q1;
set all_neg;
if emp1=1 then do;
      if change=1 then cell=1;
      else if change=3 then cell=2;
      else cell=3;
end;
else if emp1=2 then do;
      if change=1 then cell=4;
      else if change=3 then cell=5;
      else cell=6;
end;
else if emp1=3 then do;
      if change=1 then cell=7;
      else if change=3 then cell=8;
      else cell=9;
end;
else cell=10;
if NR_0+LR_0=0 then do;
      if NR_1=0 then do;
            if prior ge 0 then output conf1;
            else output conq1;
      end;
      else output only0;
end;
else if NR_0+LR_0=1 then do;
      if prior ge 0 then output for1;
      else output forq1;
end;
else output Q1;
run;

*calculate current link relatives;
*preliminary;
*combine PRs that reported in month t-1;
data PR;
set conf1;
run;

proc sort data=PR;
by ind month;
```

```
run;

proc summary data=PR nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_PR sum=p c;
run;

data LR_PR (drop=p c);
set LR_PR (drop=_type_ _freq_);
lr_PR=c/p;
run;

proc sort data=LR_PR;
by ind month;
run;

*final;
*combine LRs that reported in month t-1 with PRs that reported in
month t-1;
data LR;
set for1;
if LR_0=1;
run;

data Rpt;
set PR LR;
run;

proc sort data=Rpt;
by ind month;
run;

proc summary data=Rpt nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_Rpt sum=p c;
run;

data LR_Rpt (drop=p c);
set LR_Rpt (drop=_type_ _freq_);
lr_R=c/p;
run;

proc sort data=LR_Rpt;
by ind month;
run;

*First level estimation;
*create variables for use in comparions for constant reporters;
*0 refers to reported values;
*1 refers to current size x prior change imputed values;
*f refers to final values;
%macro createvar(var);
p0=prior;
```

```
c0=curr;
p0p=prior;
c0p=curr;
p0f=prior;
c0f=curr;
p1=prior;
c1=curr;
p1f=prior;
c1f=curr;
%mend;

data conf1;
set conf1;
%createvar(1);
run;

*Carry out imputation;
***********;
*month t-1 Rpt;
*calcuate link relative for size x change;
proc sort data=conf1;
by ind month cell;
run;

proc summary data=conf1 nway;
by ind month cell;
var p0 c0;
weight selwt;
output out=LR_sc sum=p c;
run;

data LR_sc (drop=p c);
set LR_sc (drop=_type_ _freq_);
lr_sc=c/p;
run;

proc sort data=LR_sc;
by ind month cell;
run;

*merge with ind link relatives;
data LR_conf1;
merge LR_PR LR_Rpt LR_sc;
by ind month;
run;

*impute for missing values using emp x change, ind, model;
*missing month t, month t-1 emp>0, prior change available;
proc sort data= for1;
by ind month cell;
run;

data for1adj;
merge LR_conf1 for1 (in=a);
by ind month cell;
if a;
c1=prior*lr_sc;
```

```
p1=prior;
if LR_0=1 then do;
      p0p=prior;
      c0p=prior*lr_PR;
      p0f=prior;
      c0f=curr;
      c1f=curr;
      p1f=prior;
end;
else if NR_0=1 then do;
      c1f=c1;
      p1f=p1;
end;
run;

*Final estimation;
*use all available records;

data imp_neg;
set conf1 for1adj;
run;

********************;
*repeat for other half-sample;
*Group into tertiles by ch-within ind, month month t-2 emp group;
*yield low=0, med=1, hi=2, unk=.;
proc sort data=all_pos;
by ind month group2;
run;

proc rank data=all_pos out=all_pos groups=3;
by ind month group2;
var ch;
ranks ch_r;
run;

data all_pos (drop=ch);
set all_pos;
if emp1 le 3 then rch=ch_r;
      else rch=.;
if rch ge 0 then do;
      if rch = 0 then change=1;
      else if rch =1 then change=2;
      else if rch = 2 then change=3;
end;
else change=4;
run;

data conf1 conq1 only0
      for1 forq1
      Q1;
set all_pos;
if emp1=1 then do;
      if change=1 then cell=1;
      else if change=3 then cell=2;
      else cell=3;
end;
```

```
else if emp1=2 then do;
      if change=1 then cell=4;
      else if change=3 then cell=5;
      else cell=6;
end;
else if emp1=3 then do;
      if change=1 then cell=7;
      else if change=3 then cell=8;
      else cell=9;
end;
else cell=10;
if NR_0+LR_0=0 then do;
      if NR_1=0 then do;
            if prior ge 0 then output conf1;
            else output conq1;
      end;
      else output only0;
end;
else if NR_0+LR_0=1 then do;
      if prior ge 0 then output for1;
      else output forq1;
end;
else output Q1;
run;

*calculate current link relatives;
*preliminary;
*combine PRs that reported in month t-1;
data PR;
set conf1;
run;

proc sort data=PR;
by ind month;
run;

proc summary data=PR nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_PR sum=p c;
run;

data LR_PR (drop=p c);
set LR_PR (drop=_type_ _freq_);
lr_PR=c/p;
run;

proc sort data=LR_PR;
by ind month;
run;

*final;
*combine LRs that reported in month t-1 with PRs that reported in
month t-1;
data LR;
set for1;
```

```
if LR_0=1;
run;

data Rpt;
set PR LR;
run;

proc sort data=Rpt;
by ind month;
run;

proc summary data=Rpt nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_Rpt sum=p c;
run;

data LR_Rpt (drop=p c);
set LR_Rpt (drop=_type_ _freq_);
lr_R=c/p;
run;

proc sort data=LR_Rpt;
by ind month;
run;

*First level estimation;
*create variables for use in comparions for constant reporters;
*0 refers to reported values;
*1 refers to current size x prior change imputed values;
*f refers to final values;

data conf1;
set conf1;
%createvar(1);
run;

*Carry out imputation;
***********;
*month t-1 Rpt;
*calcuate link relative for size x change;
proc sort data=conf1;
by ind month cell;
run;

proc summary data=conf1 nway;
by ind month cell;
var p0 c0;
weight selwt;
output out=LR_sc sum=p c;
run;

data LR_sc (drop=p c);
set LR_sc (drop=_type_ _freq_);
lr_sc=c/p;
run;
```

```
proc sort data=LR_sc;
by ind month cell;
run;

*merge with ind link relatives;
data LR_conf1;
merge LR_PR LR_Rpt LR_sc;
by ind month;
run;

*impute for missing values using emp x change, ind, model;
*missing month t, month t-1 emp>0, prior change available;
proc sort data= for1;
by ind month cell;
run;

data for1adj;
merge LR_conf1 for1 (in=a);
by ind month cell;
if a;
c1=prior*lr_sc;
p1=prior;
if LR_0=1 then do;
      p0p=prior;
      c0p=prior*lr_PR;
      p0f=prior;
      c0f=curr;
      c1f=curr;
      p1f=prior;
end;
else if NR_0=1 then do;
      c1f=c1;
      p1f=p1;
end;
run;

*Final estimation;
*use all available records;

data imp_pos;
set conf1 for1adj;
run;

data imp;
set imp_neg imp_pos;
run;

proc sort data=imp;
by ind month;
run;

proc summary data=imp nway;
by ind month;
var p0 c0 p0f c0f p1 c1 p1f c1f;
weight selwt;
output out=LR_imp sum=p0 c0 p0f c0f pI cI pIf cIf;
```

```
run;

data LR_imp_&z (drop=p0 c0 p0f c0f pI cI pIf cIf);
set LR_imp (drop=_type_ _freq_);
lr0_&z=c0/p0;
lr0f_&z=c0f/p0f;
lr1_&z=cI/pI;
lr1f_&z=cIf/pIf;
run;

proc sort data=LR_imp_&z;
by ind month;
run;
%end;
%mend;

%rg(80);
run;

proc sort data=LR_imp_all;
by ind month;
run;

%macro together(rg);
merge LR_imp_all (keep=ind month lr0 lr0f lr1 lr1f)
%do a=1 %to &rg;
LR_imp_&a
%end;
;
%mend;

data LR_imp0 (keep=ind month lr0 lr0f lr0_1-lr0_80 lr0f_1-lr0f_80)
      LR_imp1 (keep=ind month lr1 lr1f lr1_1-lr1_80 lr1f_1-lr1f_80);
%together(80);
by ind month;
output LR_imp0;
output LR_imp1;
run;

data hold.LR_imp0;
set LR_imp0;
run;

data hold.LR_imp1;
set LR_imp1;
run;

data hold.LR_imp_all;
set LR_imp_all;
run;

data LR_imp0p (keep=ind month lr0 lr0_1-lr0_80)
      LR_imp0f (keep=ind month lr0f lr0f_1-lr0f_80);
set hold.LR_imp0;
output LR_imp0p;
output LR_imp0f;
run;
```

```
PROC EXPORT DATA= work.LR_imp0p
            OUTFILE= "c:\CES Data\RG0bp.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

PROC EXPORT DATA= work.LR_imp0f
            OUTFILE= "c:\CES Data\RG0bf.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

data LR_imp1p (keep=ind month lr1 lr1_1-lr1_80)
      LR_imp1f (keep=ind month lr1f lr1f_1-lr1f_80);
set hold.LR_imp1;
output LR_imp1p;
output LR_imp1f;
run;

PROC EXPORT DATA= work.LR_imp1p
            OUTFILE= "c:\CES Data\RG1bp.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

PROC EXPORT DATA= work.LR_imp1f
            OUTFILE= "c:\CES Data\RG1bf.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

*options mprint;
Libname hold "c:\CES Data";

*Program Name: x:Research Project/Paper Programs/Revisions/Final/;
*AE Estimation variance-2size-model;
*Calculates half-sample estimates;

%macro createvar(var);
p0=prior;
c0=curr;
p0p=prior;
c0p=curr;
p0f=prior;
c0f=curr;
p1=prior;
c1=curr;
p1f=prior;
c1f=curr;
p2=prior;
c2=curr;
p2f=prior;
c2f=curr;
%mend;

*Exclude month t atypicals;
*Exclude if nonreporter in month t-1;
*Create emp class (<10, 10+) based on month t-1, month t reported
employment;
*Create size groupings based on month t-2 reported emp;
```

```
*(if atyp in month t-1 then assume reported emp for month t-2
unknown);
*calculate change, relative change from month t-2 to t-1;
data all (keep=ind month size n ch selwt curr prior emp1 group2 LR_0
NR_0 NR_1 NR_2 R);
set hold.analysis1r (keep=LR_0 NR_0 NR_1 NR_2 atyp_0 atyp_1
      ind month y_0 y_1 y_2 selwt size);
if NR_1 = 1 then delete;
if atyp_0 ge 1 then delete;
if y_1 = . then delete;
n=1;
if atyp_1 ge 1 then do;
      NR_2=1;
      y_2=.;
end;
if selwt = . then selwt=1;
if y_1 le 9 then emp1=1;
      else emp1=2;
if y_2 = . then group2=0;
      else if y_2 le 9 then group2=1;
      else if y_2 le 19 then group2=2;
      else if y_2 le 49 then group2=3;
      else if y_2 le 99 then group2=4;
      else if y_2 le 249 then group2=5;
      else if y_2 le 499 then group2=6;
      else group2=7;
if atyp_1 ge 1 then ch=.;
if NR_2=0 then ch=y_1-y_2;
rename y_1=prior;
rename y_0=curr;
if LR_0=1 then R=2;
      else if NR_0=1 then R=3;
      else R=1;
run;

*Group into tertiles by ch-within ind, month month t-2 emp group;
*yield low=0, med=1, hi=2, unk=.;
proc sort data=all;
by ind month group2;
run;

proc rank data=all out=all groups=3;
by ind month group2;
var ch;
ranks ch_r;
run;

*rename change to make low=1, med=2, hi=3, unk=0;
*use actual change for month t-1 emp <10, relative change for month
t-1 emp 10+
*create dummy variables for use in model estimation;
data all (drop=ch);
set all;
if emp1 = 1 then rch=ch_r;
      else rch=.;
if rch ge 0 then do;
      if rch = 0 then change=1;
      else if rch =1 then change=2;
      else if rch = 2 then change=3;
end;
else change=4;
run;

*Create subsets for use in estimating LRs, imputation;
*conf1: LR in month t, R in month t-1 and t-2, month t-1 emp>0;
*for1: not LR in month t, R in month t-1 and t-2, month t-1 emp>0;
*q data sets should be empty;
data conf1 conq1 only0
      for1 forq1
      Q1;
set all;
if emp1=1 then do;
      if change=1 then cell=1;
      else if change=3 then cell=2;
      else cell=3;
end;
else cell=4;
if NR_0+LR_0=0 then do;
      if NR_1=0 then do;
            if prior ge 0 then output conf1;
            else output conq1;
      end;
      else output only0;
end;
else if NR_0+LR_0=1 then do;
      if prior ge 0 then output for1;
      else output forq1;
end;
else output Q1;
run;

*calculate current link relatives;
*preliminary;
*combine PRs that reported in month t-1;
data PR;
set conf1;
run;

proc sort data=PR;
by ind month;
run;

proc summary data=PR nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_PR sum=p c;
run;

data LR_PR (drop=p c);
set LR_PR (drop=_type_ _freq_);
lr_PR=c/p;
run;
```

```
proc sort data=LR_PR;
by ind month;
run;

*final;
*combine LRs that reported in month t-1 with PRs that reported in
month t-1;
data LR;
set for1;
if LR_0=1;
run;

data Rpt;
set PR LR;
run;

proc sort data=Rpt;
by ind month;
run;

proc summary data=Rpt nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_Rpt sum=p c;
run;

data LR_Rpt (drop=p c);
set LR_Rpt (drop=_type_ _freq_);
lr_R=c/p;
run;

proc sort data=LR_Rpt;
by ind month;
run;

*First level estimation;
*create variables for use in comparions for constant reporters;
*0 refers to reported values;
*1 refers to post-stratification imputed values;
*2 refers to model imputed values;
*f refers to final values;
data conf1;
set conf1;
%createvar(1);
run;

*Carry out imputation;
***********;
*month t-1 emp>0, prior change available;
*calcuate link relative;
proc sort data=conf1;
by ind month cell;
run;

proc summary data=conf1 nway;
by ind month cell;
```

```
var p0 c0;
id emp1;
weight selwt;
output out=LR_sc sum=p c;
run;

data LR_sc (drop=p c);
set LR_sc (drop=_type_ _freq_);
lr_sc=c/p;
run;

proc sort data=LR_sc;
by ind month emp1;
run;

*create records for use in model;
*reference value is for change=2;
data LR_model;
set LR_sc;
if cell=3 or cell=4;
rename lr_sc=lr_model;
run;

proc sort data=LR_model;
by ind month emp1;
run;

*Read file of factors for model;
data factors;
infile 'c:\CES Data\Paper\Model Parameters.csv' delimiter=',';
input month ind emp1 pC1 pC2 pC3;
run;

proc sort data=factors;
by ind month emp1;
run;

*merge model factors with emp x change link relatives;
data LR_sc_model;
merge LR_model factors LR_sc;
by ind month emp1;
run;

*merge with ind link relatives;
data LR_all;
merge LR_PR LR_Rpt LR_sc_model;
by ind month;
run;

proc sort data=LR_all;
by ind month cell;
run;

*impute for missing values using emp x change, ind, model;
*missing month t, month t-1 emp>0, prior change available;
proc sort data= for1;
by ind month cell;
```

```
run;

proc print data=LR_all;
run;

data for1adj;
merge LR_all for1 (in=a);
by ind month cell;
if a;
c1=prior*lr_sc;
p1=prior;
if cell=1 then c2=prior*(lr_model + pC1);
else if cell=2 then c2=prior*(lr_model + pC2);
else c2=prior*lr_model;
p2=prior;
if LR_0=1 then do;
      p0p=prior;
      c0p=prior*lr_PR;
      p0f=prior;
      c0f=curr;
      c1f=curr;
      p1f=prior;
      c2f=curr;
      p2f=prior;
end;
else if NR_0=1 then do;
      c1f=c1;
      p1f=p1;
      c2f=c2;
      p2f=p2;
end;
run;

*Final estimation;
*use all available records;
data imp;
set conf1 for1adj;
run;

proc sort data=imp;
by ind month;
run;

proc summary data=imp nway;
by ind month;
var p0 c0 p0f c0f p1 c1 p1f c1f p2 c2 p2f c2f;
weight selwt;
output out=LR_imp sum=p0 c0 p0f c0f pI cI pIf cIf pM cM pMf cMf;
run;

data LR_imp_all (drop=p0 c0 p0f c0f pI cI pIf cIf pM cM pMf cMf);
set LR_imp (drop=_type_ _freq_);
lr0=c0/p0;
lr0f=c0f/p0f;
lr1=cI/pI;
lr1f=cIf/pIf;
lr2=cM/pM;
```

```
lr2f=cMf/pMf;
run;

proc sort data=LR_imp_all;
by ind month;
run;


proc sort data=imp_LR;
by ind month cell;
run;

proc summary data=imp_LR nway;
by ind month cell;
var p0 c0 p0p c0p p0f c0f p1 c1 p1f c1f p2 c2 p2f c2f;
weight selwt;
output out=LR_imp sum=p0 c0 p0p c0p p0f c0f pI cI pIf cIf pM cM pMf
cMf;
run;

data LR_imp_cell (drop=p0 c0 p0p c0p p0f c0f pI cI pIf cIf pM cM pMf
cMf);
set LR_imp (drop=_type_ _freq_);
lr0p=c0p/p0p;
lr0f=c0f/p0f;
lr1a=cI/pI;
lr2=cM/pM;
run;

proc sort data=LR_imp_cell;
by ind month cell;
run;


PROC EXPORT DATA= WORK.LR_imp_cell
            OUTFILE= "c:\CES Data\Revisions\Final\Link
Relatives_2size_cell.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;


%macro rg(rg);
%do z=1 %to &rg;
data all_neg (keep=ind month size n ch selwt curr prior emp1 group2
LR_0 NR_0 NR_1 NR_2)
      all_pos (keep=ind month size n ch selwt curr prior emp1 group2
LR_0 NR_0 NR_1 NR_2);
set hold.analysis1r (keep=LR_0 NR_0 NR_1 NR_2 atyp_0 atyp_1 h&z
      ind month y_0 y_1 y_2 selwt size);
if NR_1 = 1 then delete;
if atyp_0 ge 1 then delete;
if h&z=. then delete;
if y_1 = . then delete;
n=1;
if atyp_1 ge 1 then do;
      NR_2=1;
      y_2=.;
```

```
end;
if selwt = . then selwt=1;
selwt=1+0.5*h&z;
if y_1 le 9 then emp1=1;
      else emp1=2;
if y_2 = . then group2=0;
      else if y_2 le 9 then group2=1;
      else if y_2 le 19 then group2=2;
      else if y_2 le 49 then group2=3;
      else if y_2 le 99 then group2=4;
      else if y_2 le 249 then group2=5;
      else if y_2 le 499 then group2=6;
      else group2=7;
if atyp_1 ge 1 then     ch=.;
if NR_2=0 then ch=y_1-y_2;
rename y_1=prior;
rename y_0=curr;
if h&z=-1 then output all_neg;
else if h&z=1 then output all_pos;
run;

****************;
*run for one half sample;
*Group into tertiles by ch, relch-within ind, month month t-2 emp
group;
*yield low=0, med=1, hi=2, unk=.;
proc sort data=all_neg;
by ind month group2;
run;

proc rank data=all_neg out=all_neg groups=3;
by ind month group2;
var ch;
ranks ch_r;
run;

data all_neg (drop=ch);
set all_neg;
if emp1 = 1 then rch=ch_r;
      else rch=.;
if rch ge 0 then do;
      if rch = 0 then change=1;
      else if rch =1 then change=2;
      else if rch = 2 then change=3;
end;
else change=4;
run;

data conf1 conq1 only0
      for1 forq1
      Q1;
set all_neg;
if emp1=1 then do;
      if change=1 then cell=1;
      else if change=3 then cell=2;
      else cell=3;
end;
```

```
else cell=4;
if NR_0+LR_0=0 then do;
      if NR_1=0 then do;
            if prior ge 0 then output conf1;
            else output conq1;
      end;
      else output only0;
end;
else if NR_0+LR_0=1 then do;
      if prior ge 0 then output for1;
      else output forq1;
end;
else output Q1;
run;

*calculate current link relatives;
*preliminary;
*combine PRs that reported in month t-1;
data PR;
set conf1;
run;

proc sort data=PR;
by ind month;
run;

proc summary data=PR nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_PR sum=p c;
run;

data LR_PR (drop=p c);
set LR_PR (drop=_type_ _freq_);
lr_PR=c/p;
run;

proc sort data=LR_PR;
by ind month;
run;

*final;
*combine LRs that reported in month t-1 with PRs that reported in
month t-1;
data LR;
set for1;
if LR_0=1;
run;

data Rpt;
set PR LR;
run;

proc sort data=Rpt;
by ind month;
run;
```

```
proc summary data=Rpt nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_Rpt sum=p c;
run;

data LR_Rpt (drop=p c);
set LR_Rpt (drop=_type_ _freq_);
lr_R=c/p;
run;

proc sort data=LR_Rpt;
by ind month;
run;

*First level estimation;
*create variables for use in comparions for constant reporters;
*0 refers to reported values;
*1 refers to post-stratification imputed values;
*2 refers to model imputed values;
*f refers to final values;
data conf1;
set conf1;
%createvar(1);
run;

*Carry out imputation;
***********;
*month t-1 emp>0, prior change available;
*calcuate link relative;
proc sort data=conf1;
by ind month cell;
run;

proc summary data=conf1 nway;
by ind month cell;
var p0 c0;
id emp1;
weight selwt;
output out=LR_sc sum=p c;
run;

data LR_sc (drop=p c);
set LR_sc (drop=_type_ _freq_);
lr_sc=c/p;
run;

proc sort data=LR_sc;
by ind month emp1;
run;

*create records for use in model;
*reference value is for change=2;
data LR_model;
set LR_sc;
```

```
if cell=3 or cell=4;
rename lr_sc=lr_model;
run;

proc sort data=LR_model;
by ind month emp1;
run;

*Read file of factors for model;
data factors;
infile 'c:\CES Data\Paper\Model Parameters.csv' delimiter=',';
input month ind emp1 pC1 pC2 pC3;
run;

proc sort data=factors;
by ind month emp1;
run;

*merge model factors with emp x change link relatives;
data LR_sc_model;
merge LR_model factors LR_sc;
by ind month emp1;
run;

*merge with ind link relatives;
data LR_all;
merge LR_PR LR_Rpt LR_sc_model;
by ind month;
run;

proc sort data=LR_all;
by ind month cell;
run;

*impute for missing values using emp x change, ind, model;
*missing month t, month t-1 emp>0, prior change available;
proc sort data= for1;
by ind month cell;
run;

data for1adj;
merge LR_all for1 (in=a);
by ind month cell;
if a;
c1=prior*lr_sc;
p1=prior;
if cell=1 then c2=prior*(lr_model + pC1);
else if cell=2 then c2=prior*(lr_model + pC2);
else c2=prior*lr_model;
p2=prior;
if LR_0=1 then do;
      p0p=prior;
      c0p=prior*lr_PR;
      p0f=prior;
      c0f=curr;
      c1f=curr;
      p1f=prior;
      c2f=curr;
      p2f=prior;
end;
else if NR_0=1 then do;
      c1f=c1;
      p1f=p1;
      c2f=c2;
      p2f=p2;
end;
run;

*Final estimation;
*use all available records;
data imp_neg;
set conf1 for1adj;
run;


********************;
*repeat for other half-sample;
*Group into tertiles by ch-within ind, month month t-2 emp group;
*yield low=0, med=1, hi=2, unk=.;

proc sort data=all_pos;
by ind month group2;
run;

proc rank data=all_pos out=all_pos groups=3;
by ind month group2;
var ch;
ranks ch_r;
run;

data all_pos (drop=ch);
set all_pos;
if emp1 = 1 then rch=ch_r;
      else rch=.;
if rch ge 0 then do;
      if rch = 0 then change=1;
      else if rch =1 then change=2;
      else if rch = 2 then change=3;
end;
else change=4;
run;

data conf1 conq1 only0
      for1 forq1
      Q1;
set all_pos;
if emp1=1 then do;
      if change=1 then cell=1;
      else if change=3 then cell=2;
      else cell=3;
end;
else cell=4;
if NR_0+LR_0=0 then do;
      if NR_1=0 then do;
```

```
            if prior ge 0 then output conf1;
            else output conq1;
      end;
      else output only0;
end;
else if NR_0+LR_0=1 then do;
      if prior ge 0 then output for1;
      else output forq1;
end;
else output Q1;
run;

*calculate current link relatives;
*preliminary;
*combine PRs that reported in month t-1;
data PR;
set conf1;
run;

proc sort data=PR;
by ind month;
run;

proc summary data=PR nway;
by ind month;
var prior curr;
weight selwt;
output out=LR_PR sum=p c;
run;

data LR_PR (drop=p c);
set LR_PR (drop=_type_ _freq_);
lr_PR=c/p;
run;

proc sort data=LR_PR;
by ind month;
run;

*final;
*combine LRs that reported in month t-1 with PRs that reported in
month t-1;
data LR;
set for1;
if LR_0=1;
run;

data Rpt;
set PR LR;
run;

proc sort data=Rpt;
by ind month;
run;

proc summary data=Rpt nway;
by ind month;
```

```
var prior curr;
weight selwt;
output out=LR_Rpt sum=p c;
run;

data LR_Rpt (drop=p c);
set LR_Rpt (drop=_type_ _freq_);
lr_R=c/p;
run;

proc sort data=LR_Rpt;
by ind month;
run;

*First level estimation;
*create variables for use in comparions for constant reporters;
*0 refers to reported values;
*1 refers to post-stratification imputed values;
*2 refers to model imputed values;
*f refers to final values;
data conf1;
set conf1;
%createvar(1);
run;

*Carry out imputation;
***********;
*month t-1 emp>0, prior change available;
*calcuate link relative;
proc sort data=conf1;
by ind month cell;
run;

proc summary data=conf1 nway;
by ind month cell;
var p0 c0;
id emp1;
weight selwt;
output out=LR_sc sum=p c;
run;

data LR_sc (drop=p c);
set LR_sc (drop=_type_ _freq_);
lr_sc=c/p;
run;

proc sort data=LR_sc;
by ind month emp1;
run;

*create records for use in model;
*reference value is for change=2;
data LR_model;
set LR_sc;
if cell=3 or cell=4;
rename lr_sc=lr_model;
run;
```

```
proc sort data=LR_model;
by ind month emp1;
run;

*Read file of factors for model;
data factors;
infile 'c:\CES Data\Paper\Model Parameters.csv' delimiter=',';
input month ind emp1 pC1 pC2 pC3;
run;

proc sort data=factors;
by ind month emp1;
run;

*merge model factors with emp x change link relatives;
data LR_sc_model;
merge LR_model factors LR_sc;
by ind month emp1;
run;

*merge with ind link relatives;
data LR_all;
merge LR_PR LR_Rpt LR_sc_model;
by ind month;
run;

proc sort data=LR_all;
by ind month cell;
run;

*impute for missing values using emp x change, ind, model;
*missing month t, month t-1 emp>0, prior change available;
proc sort data= for1;
by ind month cell;
run;

data for1adj;
merge LR_all for1 (in=a);
by ind month cell;
if a;
c1=prior*lr_sc;
p1=prior;
if cell=1 then c2=prior*(lr_model + pC1);
else if cell=2 then c2=prior*(lr_model + pC2);
else c2=prior*lr_model;
p2=prior;
if LR_0=1 then do;
      p0p=prior;
      c0p=prior*lr_PR;
      p0f=prior;
      c0f=curr;
      c1f=curr;
      p1f=prior;
      c2f=curr;
      p2f=prior;
end;
```

```
else if NR_0=1 then do;
      c1f=c1;
      p1f=p1;
      c2f=c2;
      p2f=p2;
end;
run;

*Final estimation;
*use all available records;
data imp_pos;
set conf1 for1adj;
run;

data imp;
set imp_neg imp_pos;
run;

proc sort data=imp;
by ind month;
run;

proc summary data=imp nway;
by ind month;
var p0 c0 p0f c0f p1 c1 p1f c1f p2 c2 p2f c2f;
weight selwt;
output out=LR_imp sum=p0 c0 p0f c0f pI cI pIf cIf pM cM pMf cMf;
run;

data LR_imp_&z (drop=p0 c0 p0f c0f pI cI pIf cIf pM cM pMf cMf);
set LR_imp (drop=_type_ _freq_);
lr0_&z=c0/p0;
lr0f_&z=c0f/p0f;
lr1_&z=cI/pI;
lr1f_&z=cIf/pIf;
lr2_&z=cM/pM;
lr2f_&z=cMf/pMf;
run;

proc sort data=LR_imp_&z;
by ind month;
run;

%end;
%mend;

%rg(80);
run;

proc sort data=LR_imp_all;
by ind month;
run;

%macro together(rg);
merge LR_imp_all (keep=ind month lr0 lr0f lr1 lr1f lr2 lr2f)
%do a=1 %to &rg;
LR_imp_&a
```

```
%end;
;
%mend;

data LR_imp0 (keep=ind month lr0 lr0f lr0_1-lr0_80 lr0f_1-lr0f_80)
      LR_imp1 (keep=ind month lr1 lr1f lr1_1-lr1_80 lr1f_1-lr1f_80)
      LR_imp2 (keep=ind month lr2 lr2f lr2_1-lr2_80 lr2f_1-lr2f_80);
%together(80);
by ind month;
output LR_imp0;
output LR_imp1;
output LR_imp2;
run;

proc print data=LR_imp0;
run;

data hold.LR_imp0;
set LR_imp0;
run;

data hold.LR_imp1;
set LR_imp1;
run;

data hold.LR_imp2;
set LR_imp2;
run;

data hold.LR_imp_all;
set LR_imp_all;
run;

data LR_imp0p (keep=ind month lr0 lr0_1-lr0_80)
      LR_imp0f (keep=ind month lr0f lr0f_1-lr0f_80);
set hold.LR_imp0;
output LR_imp0p;
output LR_imp0f;
run;

PROC EXPORT DATA= work.LR_imp0p
            OUTFILE= "c:\CES Data\RG0ap.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

PROC EXPORT DATA= work.LR_imp0f
            OUTFILE= "c:\CES Data\RG0af.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

data LR_imp1p (keep=ind month lr1 lr1_1-lr1_80)
      LR_imp1f (keep=ind month lr1f lr1f_1-lr1f_80);
set hold.LR_imp1;
output LR_imp1p;
output LR_imp1f;
run;
```

```
PROC EXPORT DATA= work.LR_imp1p
            OUTFILE= "c:\CES Data\RG1ap.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

PROC EXPORT DATA= work.LR_imp1f
            OUTFILE= "c:\CES Data\RG1af.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

data LR_imp2p (keep=ind month lr2 lr2_1-lr2_80)
      LR_imp2f (keep=ind month lr2f lr2f_1-lr2f_80);
set hold.LR_imp2;
output LR_imp2p;
output LR_imp2f;
run;

PROC EXPORT DATA= work.LR_imp2p
            OUTFILE= "c:\CES Data\RG2p.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

PROC EXPORT DATA= work.LR_imp2f
            OUTFILE= "c:\CES Data\RG2f.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;
```

I. Estimated Link Relatives

Estimated Link Relatives: Construction
March 2000 - December 2002

	Current					Model 1A					Model 1B					Model 2				
	Preliminary		Final			Preliminary		Final			Preliminary		Final			Preliminary		Final		
Month	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision
Mar-00	1.0259	0.0062	1.0268	0.0038	0.0009	1.0264	0.0057	1.0270	0.0038	0.0006	1.0262	0.0057	1.0270	0.0038	0.0008					
Apr-00	1.0333	0.0049	1.0342	0.0041	0.0009	1.0335	0.0047	1.0344	0.0040	0.0009	1.0334	0.0046	1.0344	0.0040	0.0010					
May-00	1.0244	0.0160	1.0233	0.0144	-0.0011	1.0244	0.0159	1.0236	0.0142	-0.0009	1.0245	0.0157	1.0235	0.0142	-0.0009					
Jun-00	1.0236	0.0251	1.0250	0.0234	0.0015	1.0236	0.0251	1.0249	0.0236	0.0013	1.0236	0.0252	1.0250	0.0235	0.0014					
Jul-00	1.0099	0.0076	1.0083	0.0089	-0.0016	1.0083	0.0105	1.0075	0.0106	-0.0008	1.0066	0.0134	1.0059	0.0135	-0.0008					
Aug-00	1.0039	0.0058	1.0038	0.0035	-0.0001	1.0040	0.0058	1.0039	0.0036	-0.0001	1.0039	0.0059	1.0038	0.0039	-0.0001					
Sep-00	0.9911	0.0206	0.9927	0.0228	0.0017	0.9910	0.0207	0.9926	0.0227	0.0016	0.9913	0.0202	0.9927	0.0225	0.0014					
Oct-00	1.0085	0.0087	1.0067	0.0071	-0.0018	1.0085	0.0088	1.0068	0.0071	-0.0017	1.0086	0.0090	1.0067	0.0071	-0.0019					
Nov-00	0.9796	0.0038	0.9780	0.0037	-0.0016	0.9796	0.0038	0.9781	0.0037	-0.0015	0.9796	0.0038	0.9781	0.0037	-0.0014					
Dec-00	0.9661	0.0076	0.9661	0.0067	0.0001	0.9660	0.0077	0.9662	0.0067	0.0002	0.9656	0.0082	0.9662	0.0068	0.0005					
Jan-01	0.9684	0.0068	0.9658	0.0066	-0.0025	0.9683	0.0069	0.9662	0.0061	-0.0021	0.9685	0.0071	0.9663	0.0061	-0.0022					
Feb-01	0.9961	0.0040	0.9969	0.0036	0.0008	0.9958	0.0039	0.9967	0.0036	0.0009	0.9959	0.0039	0.9967	0.0035	0.0008					
Mar-01	1.0092	0.0076	1.0097	0.0088	0.0005	1.0092	0.0075	1.0097	0.0086	0.0005	1.0090	0.0075	1.0097	0.0086	0.0006					
Apr-01	1.0279	0.0116	1.0277	0.0095	-0.0002	1.0278	0.0116	1.0277	0.0095	-0.0001	1.0276	0.0118	1.0276	0.0095	0.0000	1.0282	0.0111	1.0279	0.0094	-0.0003
May-01	1.0314	0.0084	1.0323	0.0164	0.0009	1.0311	0.0083	1.0322	0.0158	0.0011	1.0309	0.0086	1.0321	0.0159	0.0012	1.0316	0.0082	1.0323	0.0157	0.0008
Jun-01	1.0277	0.0271	1.0255	0.0296	-0.0022	1.0275	0.0267	1.0256	0.0294	-0.0019	1.0271	0.0266	1.0255	0.0294	-0.0015	1.0280	0.0257	1.0257	0.0292	-0.0022
Jul-01	1.0101	0.0043	1.0120	0.0036	0.0019	1.0103	0.0041	1.0119	0.0035	0.0016	1.0100	0.0043	1.0117	0.0035	0.0018	1.0099	0.0044	1.0118	0.0035	0.0018
Aug-01	0.9970	0.0082	0.9972	0.0038	0.0002	0.9970	0.0081	0.9972	0.0039	0.0002	0.9970	0.0083	0.9972	0.0039	0.0002	0.9973	0.0081	0.9973	0.0039	0.0000
Sep-01	0.9833	0.0129	0.9839	0.0142	0.0006	0.9832	0.0131	0.9839	0.0142	0.0007	0.9833	0.0129	0.9839	0.0141	0.0006	0.9836	0.0131	0.9840	0.0142	0.0004
Oct-01	0.9934	0.0029	0.9934	0.0033	-0.0001	0.9935	0.0028	0.9935	0.0032	0.0000	0.9934	0.0029	0.9934	0.0032	0.0000	0.9935	0.0028	0.9935	0.0032	0.0000
Nov-01	0.9764	0.0054	0.9764	0.0042	0.0000	0.9762	0.0058	0.9764	0.0044	0.0001	0.9763	0.0056	0.9764	0.0043	0.0001	0.9761	0.0055	0.9763	0.0043	0.0002
Dec-01	0.9717	0.0161	0.9702	0.0106	-0.0016	0.9716	0.0161	0.9703	0.0109	-0.0013	0.9718	0.0166	0.9703	0.0110	-0.0015	0.9717	0.0162	0.9703	0.0109	-0.0014
Jan-02	0.9508	0.0082	0.9531	0.0130	0.0024	0.9506	0.0081	0.9532	0.0129	0.0025	0.9505	0.0081	0.9531	0.0130	0.0027	0.9507	0.0080	0.9532	0.0130	0.0025
Feb-02	0.9866	0.0250	0.9848	0.0268	-0.0017	0.9861	0.0256	0.9847	0.0269	-0.0014	0.9861	0.0261	0.9847	0.0272	-0.0013	0.9856	0.0254	0.9845	0.0268	-0.0011
Mar-02	1.0079	0.0095	1.0089	0.0066	0.0010	1.0076	0.0098	1.0087	0.0069	0.0011	1.0074	0.0098	1.0086	0.0070	0.0012	1.0072	0.0099	1.0086	0.0070	0.0015
Apr-02	1.0261	0.0081	1.0256	0.0086	-0.0005	1.0258	0.0082	1.0254	0.0086	-0.0003	1.0256	0.0081	1.0252	0.0086	-0.0003					
May-02	1.0282	0.0188	1.0269	0.0203	-0.0013	1.0279	0.0189	1.0270	0.0204	-0.0009	1.0277	0.0194	1.0270	0.0208	-0.0008					
Jun-02	1.0256	0.0135	1.0241	0.0153	-0.0015	1.0255	0.0137	1.0243	0.0154	-0.0013	1.0253	0.0140	1.0243	0.0155	-0.0010					
Jul-02	1.0085	0.0040	1.0091	0.0035	0.0006	1.0085	0.0040	1.0091	0.0035	0.0006	1.0083	0.0040	1.0089	0.0034	0.0006					
Aug-02	0.9981	0.0151	0.9990	0.0176	0.0009	0.9979	0.0152	0.9990	0.0174	0.0011	0.9978	0.0152	0.9990	0.0174	0.0012					
Sep-02	0.9849	0.0119	0.9866	0.0124	0.0017	0.9848	0.0119	0.9865	0.0123	0.0017	0.9849	0.0118	0.9866	0.0122	0.0017					
Oct-02	0.9876	0.0101	0.9873	0.0109	-0.0003	0.9877	0.0099	0.9874	0.0107	-0.0002	0.9876	0.0100	0.9875	0.0107	-0.0002					
Nov-02	0.9816	0.0085	0.9827	0.0120	0.0011	0.9815	0.0083	0.9827	0.0118	0.0012	0.9813	0.0083	0.9827	0.0118	0.0014					
Dec-02	0.9671	0.0044	0.9677	0.0034	0.0006	0.9670	0.0043	0.9676	0.0034	0.0006	0.9669	0.0042	0.9676	0.0034	0.0007					
Average		0.0106		0.0107	0.0000		0.0106		0.0107	0.0001		0.0108		0.0108	0.0002		0.0115		0.0118	0.0002
Average Absolute					0.0011					0.0010					0.0010					0.0010

Estimated Link Relatives: Manufacturing
March 2000 - December 2002

	Current					Model 1A					Model 1B					Model 2				
	Preliminary		Final			Preliminary		Final			Preliminary		Final			Preliminary		Final		
Month	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision
Mar-00	1.0006	0.0036	0.9984	0.0007	-0.0022	1.0006	0.0036	0.9985	0.0005	-0.0021	1.0006	0.0036	0.9985	0.0005	-0.0021					
Apr-00	0.9998	0.0016	0.9993	0.0009	-0.0005	0.9998	0.0017	0.9993	0.0009	-0.0005	0.9998	0.0016	0.9993	0.0009	-0.0005					
May-00	1.0004	0.0009	1.0004	0.0007	0.0000	1.0004	0.0009	1.0004	0.0007	0.0000	1.0004	0.0009	1.0004	0.0007	0.0000					
Jun-00	1.0068	0.0019	1.0072	0.0022	0.0004	1.0068	0.0020	1.0071	0.0022	0.0004	1.0067	0.0020	1.0071	0.0022	0.0004					
Jul-00	0.9937	0.0043	0.9941	0.0041	0.0004	0.9936	0.0044	0.9941	0.0041	0.0005	0.9936	0.0044	0.9941	0.0041	0.0005					
Aug-00	1.0036	0.0048	1.0031	0.0033	-0.0005	1.0036	0.0048	1.0031	0.0034	-0.0005	1.0034	0.0045	1.0031	0.0034	-0.0003					
Sep-00	0.9953	0.0014	0.9978	0.0058	0.0025	0.9953	0.0014	0.9976	0.0054	0.0023	0.9952	0.0014	0.9976	0.0054	0.0023					
Oct-00	0.9974	0.0051	0.9977	0.0035	0.0004	0.9973	0.0050	0.9977	0.0036	0.0004	0.9973	0.0050	0.9977	0.0036	0.0004					
Nov-00	0.9975	0.0017	0.9973	0.0016	-0.0003	0.9975	0.0017	0.9973	0.0016	-0.0002	0.9975	0.0017	0.9973	0.0016	-0.0002					
Dec-00	0.9959	0.0018	0.9964	0.0043	0.0005	0.9959	0.0019	0.9963	0.0042	0.0004	0.9959	0.0019	0.9964	0.0041	0.0005					
Jan-01	0.9889	0.0024	0.9893	0.0037	0.0004	0.9889	0.0024	0.9893	0.0036	0.0004	0.9890	0.0024	0.9893	0.0035	0.0003					
Feb-01	0.9945	0.0039	0.9943	0.0058	-0.0002	0.9945	0.0039	0.9943	0.0057	-0.0002	0.9946	0.0039	0.9943	0.0057	-0.0002					
Mar-01	0.9962	0.0035	0.9946	0.0017	-0.0016	0.9962	0.0034	0.9947	0.0015	-0.0015	0.9961	0.0033	0.9947	0.0016	-0.0014					
Apr-01	0.9955	0.0017	0.9964	0.0026	0.0009	0.9955	0.0017	0.9963	0.0025	0.0009	0.9954	0.0017	0.9963	0.0025	0.0009	0.9954	0.0016	0.9963	0.0025	0.0009
May-01	0.9926	0.0034	0.9932	0.0038	0.0006	0.9925	0.0035	0.9931	0.0038	0.0006	0.9924	0.0037	0.9931	0.0038	0.0007	0.9925	0.0037	0.9931	0.0038	0.0006
Jun-01	1.0014	0.0023	1.0010	0.0022	-0.0004	1.0014	0.0023	1.0010	0.0022	-0.0003	1.0014	0.0022	1.0010	0.0022	-0.0004	1.0014	0.0022	1.0010	0.0022	-0.0003
Jul-01	0.9897	0.0026	0.9922	0.0023	0.0025	0.9897	0.0026	0.9921	0.0023	0.0024	0.9896	0.0026	0.9921	0.0023	0.0024	0.9897	0.0026	0.9921	0.0023	0.0024
Aug-01	0.9988	0.0071	0.9981	0.0062	-0.0008	0.9988	0.0071	0.9981	0.0062	-0.0007	0.9988	0.0071	0.9981	0.0062	-0.0007	0.9988	0.0071	0.9981	0.0062	-0.0007
Sep-01	0.9920	0.0036	0.9910	0.0024	-0.0010	0.9920	0.0036	0.9911	0.0021	-0.0009	0.9920	0.0036	0.9911	0.0021	-0.0009	0.9920	0.0036	0.9911	0.0021	-0.0010
Oct-01	0.9904	0.0014	0.9898	0.0023	-0.0007	0.9904	0.0014	0.9898	0.0022	-0.0006	0.9904	0.0014	0.9898	0.0022	-0.0006	0.9905	0.0014	0.9898	0.0022	-0.0006
Nov-01	0.9873	0.0071	0.9855	0.0097	-0.0018	0.9872	0.0072	0.9856	0.0096	-0.0016	0.9872	0.0072	0.9856	0.0096	-0.0016	0.9873	0.0072	0.9856	0.0096	-0.0017
Dec-01	0.9926	0.0030	0.9938	0.0028	0.0012	0.9926	0.0030	0.9937	0.0028	0.0011	0.9926	0.0030	0.9937	0.0028	0.0011	0.9926	0.0030	0.9937	0.0028	0.0011
Jan-02	0.9902	0.0032	0.9896	0.0049	-0.0006	0.9902	0.0032	0.9897	0.0047	-0.0006	0.9903	0.0033	0.9897	0.0047	-0.0006	0.9903	0.0033	0.9897	0.0047	-0.0006
Feb-02	0.9978	0.0036	0.9969	0.0028	-0.0010	0.9978	0.0036	0.9969	0.0028	-0.0009	0.9977	0.0034	0.9969	0.0028	-0.0008	0.9978	0.0034	0.9969	0.0028	-0.0009
Mar-02	0.9949	0.0011	0.9941	0.0015	-0.0008	0.9949	0.0012	0.9941	0.0015	-0.0007	0.9948	0.0013	0.9941	0.0015	-0.0006	0.9948	0.0014	0.9941	0.0015	-0.0007
Apr-02	0.9999	0.0047	0.9989	0.0039	-0.0010	0.9999	0.0047	0.9990	0.0040	-0.0009	0.9998	0.0046	0.9990	0.0039	-0.0008					
May-02	0.9976	0.0016	0.9968	0.0011	-0.0008	0.9975	0.0016	0.9968	0.0011	-0.0007	0.9974	0.0016	0.9968	0.0011	-0.0006					
Jun-02	1.0069	0.0017	1.0053	0.0015	-0.0006	1.0069	0.0017	1.0064	0.0015	-0.0006	1.0069	0.0017	1.0063	0.0015	-0.0005					
Jul-02	0.9957	0.0035	0.9943	0.0034	-0.0015	0.9957	0.0035	0.9944	0.0034	-0.0013	0.9957	0.0035	0.9944	0.0034	-0.0013					
Aug-02	1.0083	0.0053	1.0083	0.0083	0.0000	1.0083	0.0053	1.0083	0.0082	0.0000	1.0084	0.0056	1.0083	0.0082	-0.0001					
Sep-02	0.9938	0.0033	0.9942	0.0024	0.0004	0.9937	0.0033	0.9942	0.0024	0.0004	0.9937	0.0032	0.9941	0.0024	0.0005					
Oct-02	0.9946	0.0071	0.9944	0.0023	-0.0002	0.9945	0.0069	0.9945	0.0025	-0.0001	0.9945	0.0069	0.9945	0.0025	0.0000					
Nov-02	0.9945	0.0022	0.9929	0.0020	-0.0016	0.9945	0.0022	0.9929	0.0019	-0.0015	0.9944	0.0022	0.9930	0.0019	-0.0015					
Dec-02	0.9955	0.0043	0.9939	0.0067	-0.0016	0.9955	0.0043	0.9940	0.0066	-0.0015	0.9955	0.0042	0.9940	0.0066	-0.0015					
Average		0.0033		0.0033	-0.0003		0.0033		0.0033	-0.0002		0.0033		0.0033	-0.0002		0.0034		0.0036	-0.0001
Average Absolute					0.0009					0.0008					0.0008					0.0010

Estimated Link Relatives: Mining
March 2000 - December 2002

	Current					Model 1A					Model 1B					Model2				
	Preliminary		Final			Preliminary		Final			Preliminary		Final			Preliminary		Final		
Month	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision
Mar-00	0.9912	0.0310	0.9920	0.0255	0.0007	0.9916	0.0301	0.9921	0.0255	0.0006	0.9916	0.0292	0.9920	0.0255	0.0005					
Apr-00	1.0085	0.0103	1.0094	0.0048	0.0009	1.0091	0.0091	1.0093	0.0050	0.0002	1.0088	0.0089	1.0092	0.0051	0.0005					
May-00	1.0151	0.0217	1.0140	0.0139	-0.0011	1.0141	0.0241	1.0141	0.0143	0.0000	1.0136	0.0250	1.0141	0.0143	0.0005					
Jun-00	1.0230	0.0276	1.0193	0.0207	-0.0037	1.0222	0.0264	1.0198	0.0215	-0.0024	1.0228	0.0276	1.0196	0.0211	-0.0031					
Jul-00	1.0152	0.0253	1.0130	0.0242	-0.0022	1.0187	0.0323	1.0134	0.0248	-0.0053	1.0187	0.0323	1.0133	0.0246	-0.0054					
Aug-00	1.0149	0.0124	1.0070	0.0095	-0.0079	1.0152	0.0128	1.0078	0.0099	-0.0073	1.0112	0.0127	1.0074	0.0095	-0.0038					
Sep-00	0.9942	0.0076	0.9974	0.0086	0.0032	0.9943	0.0080	0.9973	0.0083	0.0030	0.9944	0.0086	0.9973	0.0083	0.0029					
Oct-00	0.9964	0.0082	0.9973	0.0050	0.0009	0.9963	0.0084	0.9972	0.0053	0.0010	0.9969	0.0075	0.9974	0.0051	0.0004					
Nov-00	0.9873	0.0064	0.9880	0.0028	0.0007	0.9870	0.0070	0.9878	0.0031	0.0008	0.9860	0.0091	0.9877	0.0032	0.0017					
Dec-00	0.9739	0.0180	0.9818	0.0226	0.0078	0.9740	0.0180	0.9812	0.0209	0.0072	0.9726	0.0218	0.9810	0.0220	0.0084					
Jan-01	0.9864	0.0113	0.9943	0.0231	0.0079	0.9865	0.0114	0.9939	0.0227	0.0074	0.9862	0.0095	0.9938	0.0221	0.0076					
Feb-01	0.9981	0.0167	0.9968	0.0159	-0.0014	0.9982	0.0146	0.9970	0.0157	-0.0013	0.9995	0.0131	0.9972	0.0154	-0.0023					
Mar-01	0.9976	0.0132	0.9992	0.0061	0.0016	0.9971	0.0133	0.9992	0.0062	0.0021	0.9968	0.0144	0.9993	0.0062	0.0025					
Apr-01	1.0072	0.0058	1.0071	0.0039	-0.0001	1.0075	0.0049	1.0070	0.0041	-0.0005	1.0081	0.0035	1.0070	0.0042	-0.0011	1.0073	0.0033	1.0070	0.0039	-0.0003
May-01	1.0173	0.0063	1.0173	0.0081	0.0001	1.0169	0.0061	1.0172	0.0074	0.0004	1.0178	0.0043	1.0175	0.0080	-0.0004	1.0166	0.0048	1.0172	0.0080	0.0006
Jun-01	1.0217	0.0066	1.0233	0.0060	0.0016	1.0215	0.0066	1.0233	0.0062	0.0018	1.0221	0.0077	1.0234	0.0065	0.0013	1.0214	0.0083	1.0232	0.0065	0.0018
Jul-01	1.0024	0.0168	1.0003	0.0091	-0.0021	1.0023	0.0167	1.0003	0.0090	-0.0020	1.0034	0.0152	1.0003	0.0090	-0.0031	1.0026	0.0150	1.0004	0.0090	-0.0022
Aug-01	0.9886	0.0173	0.9951	0.0158	0.0065	0.9878	0.0187	0.9949	0.0158	0.0071	0.9886	0.0174	0.9949	0.0159	0.0063	0.9891	0.0174	0.9950	0.0159	0.0059
Sep-01	1.0046	0.0236	1.0037	0.0095	-0.0009	1.0043	0.0233	1.0036	0.0097	-0.0006	1.0046	0.0249	1.0038	0.0101	-0.0009	1.0042	0.0244	1.0036	0.0099	-0.0005
Oct-01	0.9954	0.0019	0.9989	0.0025	0.0035	0.9953	0.0020	0.9986	0.0023	0.0033	0.9948	0.0025	0.9984	0.0020	0.0036	0.9955	0.0028	0.9987	0.0021	0.0032
Nov-01	0.9898	0.0050	0.9897	0.0037	-0.0001	0.9894	0.0056	0.9897	0.0038	0.0003	0.9901	0.0046	0.9900	0.0034	-0.0001	0.9895	0.0045	0.9898	0.0034	0.0003
Dec-01	0.9679	0.0104	0.9766	0.0092	0.0087	0.9678	0.0105	0.9764	0.0091	0.0086	0.9667	0.0125	0.9765	0.0089	0.0098	0.9679	0.0117	0.9764	0.0089	0.0085
Jan-02	0.9847	0.0209	0.9860	0.0096	0.0013	0.9833	0.0184	0.9859	0.0105	0.0027	0.9842	0.0168	0.9855	0.0096	0.0012	0.9831	0.0210	0.9856	0.0103	0.0026
Feb-02	0.9924	0.0060	0.9925	0.0084	0.0001	0.9924	0.0058	0.9928	0.0082	0.0004	0.9922	0.0060	0.9916	0.0068	-0.0006	0.9922	0.0045	0.9926	0.0074	0.0004
Mar-02	0.9876	0.0267	0.9836	0.0314	-0.0040	0.9874	0.0265	0.9836	0.0315	-0.0039	0.9878	0.0245	0.9836	0.0312	-0.0042	0.9870	0.0268	0.9836	0.0315	-0.0034
Apr-02	1.0091	0.0125	1.0096	0.0198	0.0004	1.0095	0.0111	1.0096	0.0184	0.0001	1.0107	0.0098	1.0107	0.0170	0.0000					
May-02	1.0201	0.0076	1.0203	0.0130	0.0002	1.0200	0.0078	1.0204	0.0126	0.0003	1.0203	0.0086	1.0207	0.0132	0.0003					
Jun-02	1.0176	0.0123	1.0155	0.0153	-0.0020	1.0176	0.0127	1.0155	0.0152	-0.0021	1.0177	0.0138	1.0156	0.0152	-0.0021					
Jul-02	0.9934	0.0076	0.9955	0.0029	0.0021	0.9931	0.0079	0.9949	0.0032	0.0019	0.9935	0.0071	0.9955	0.0027	0.0020					
Aug-02	1.0066	0.0140	0.9973	0.0248	-0.0092	1.0065	0.0140	0.9981	0.0225	-0.0084	1.0054	0.0118	0.9980	0.0226	-0.0073					
Sep-02	0.9913	0.0022	0.9949	0.0067	0.0036	0.9913	0.0021	0.9948	0.0066	0.0035	0.9917	0.0024	0.9949	0.0068	0.0032					
Oct-02	0.9891	0.0154	0.9911	0.0114	0.0020	0.9892	0.0148	0.9910	0.0116	0.0017	0.9892	0.0149	0.9910	0.0116	0.0018					
Nov-02	0.9910	0.0095	0.9818	0.0144	-0.0092	0.9908	0.0090	0.9824	0.0130	-0.0084	0.9905	0.0081	0.9824	0.0132	-0.0081					
Dec-02	0.9751	0.0459	0.9774	0.0394	0.0024	0.9751	0.0459	0.9774	0.0396	0.0023	0.9758	0.0446	0.9775	0.0395	0.0017					
Average		0.0142		0.0132	0.0004		0.0143		0.0130	0.0004		0.0141		0.0129	0.0004		0.0120		0.0097	0.0014
Average Absolute					0.0029					0.0029					0.0029					0.0026

Estimated Link Relatives: Wholesale Trade
March 2000 - December 2002

	Current					Model 1A					Model 1B					Model 2				
	Preliminary		Final			Preliminary		Final			Preliminary		Final			Preliminary		Final		
Month	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision	Link Relative	st dev	Link Relative	st dev	Revision
Mar-00	1.0044	0.0070	1.0030	0.0051	-0.0014	1.0044	0.0070	1.0031	0.0052	-0.0013	1.0044	0.0071	1.0031	0.0052	-0.0014					
Apr-00	0.9993	0.0049	1.0021	0.0052	0.0029	0.9997	0.0057	1.0019	0.0052	0.0021	0.9989	0.0044	1.0018	0.0051	0.0028					
May-00	1.0014	0.0093	1.0023	0.0036	0.0009	1.0015	0.0092	1.0023	0.0038	0.0008	1.0017	0.0092	1.0023	0.0038	0.0006					
Jun-00	1.0032	0.0121	1.0020	0.0161	-0.0012	1.0033	0.0119	1.0021	0.0157	-0.0012	1.0034	0.0118	1.0021	0.0157	-0.0012					
Jul-00	1.0021	0.0090	1.0019	0.0063	-0.0002	1.0021	0.0090	1.0019	0.0065	-0.0002	1.0021	0.0091	1.0019	0.0065	-0.0002					
Aug-00	0.9975	0.0069	0.9967	0.0088	-0.0008	0.9974	0.0070	0.9967	0.0087	-0.0007	0.9974	0.0073	0.9967	0.0089	-0.0007					
Sep-00	0.9955	0.0027	0.9951	0.0030	-0.0004	0.9954	0.0028	0.9951	0.0030	-0.0003	0.9954	0.0028	0.9950	0.0031	-0.0003					
Oct-00	1.0030	0.0021	1.0030	0.0023	0.0000	1.0028	0.0020	1.0029	0.0023	0.0002	1.0026	0.0020	1.0029	0.0023	0.0002					
Nov-00	1.0005	0.0104	1.0006	0.0059	0.0001	1.0004	0.0104	1.0006	0.0062	0.0001	1.0004	0.0106	1.0006	0.0062	0.0002					
Dec-00	0.9977	0.0118	0.9985	0.0100	0.0008	0.9977	0.0118	0.9984	0.0101	0.0007	0.9978	0.0118	0.9983	0.0103	0.0006					
Jan-01	0.9965	0.0077	0.9969	0.0071	0.0003	0.9965	0.0075	0.9969	0.0072	0.0004	0.9965	0.0076	0.9969	0.0072	0.0004					
Feb-01	0.9983	0.0027	0.9970	0.0025	-0.0012	0.9979	0.0024	0.9971	0.0024	-0.0008	0.9980	0.0024	0.9971	0.0024	-0.0009					
Mar-01	1.0004	0.0041	0.9993	0.0034	-0.0011	1.0004	0.0042	0.9994	0.0032	-0.0009	1.0005	0.0043	0.9994	0.0032	-0.0010					
Apr-01	0.9981	0.0037	0.9976	0.0020	-0.0005	0.9979	0.0036	0.9976	0.0021	-0.0003	0.9976	0.0036	0.9975	0.0021	-0.0001	0.9983	0.0034	0.9976	0.0021	-0.0007
May-01	0.9977	0.0115	0.9987	0.0122	0.0011	0.9975	0.0114	0.9986	0.0120	0.0012	0.9972	0.0118	0.9986	0.0119	0.0014	0.9981	0.0106	0.9987	0.0119	0.0007
Jun-01	1.0034	0.0038	1.0040	0.0029	0.0006	1.0034	0.0039	1.0039	0.0030	0.0006	1.0035	0.0042	1.0040	0.0031	0.0005	1.0039	0.0039	1.0041	0.0030	0.0002
Jul-01	0.9892	0.0051	0.9908	0.0045	0.0016	0.9888	0.0056	0.9906	0.0047	0.0018	0.9881	0.0065	0.9903	0.0051	0.0022	0.9894	0.0071	0.9907	0.0053	0.0013
Aug-01	0.9987	0.0105	0.9985	0.0042	-0.0002	0.9986	0.0106	0.9985	0.0044	-0.0001	0.9986	0.0106	0.9985	0.0044	-0.0001	0.9989	0.0106	0.9985	0.0044	-0.0004
Sep-01	0.9950	0.0092	0.9957	0.0100	0.0007	0.9947	0.0095	0.9956	0.0100	0.0009	0.9946	0.0096	0.9955	0.0101	0.0009	0.9951	0.0099	0.9957	0.0101	0.0006
Oct-01	0.9965	0.0076	0.9962	0.0065	-0.0002	0.9964	0.0076	0.9963	0.0066	-0.0002	0.9965	0.0077	0.9962	0.0066	-0.0002	0.9965	0.0074	0.9963	0.0066	-0.0002
Nov-01	0.9924	0.0062	0.9918	0.0070	-0.0006	0.9921	0.0059	0.9918	0.0069	-0.0003	0.9919	0.0057	0.9918	0.0069	-0.0002	0.9921	0.0056	0.9918	0.0068	-0.0003
Dec-01	1.0025	0.0032	1.0001	0.0050	-0.0023	1.0025	0.0032	1.0003	0.0049	-0.0022	1.0021	0.0030	1.0002	0.0047	-0.0019	1.0022	0.0030	1.0002	0.0048	-0.0020
Jan-02	0.9913	0.0073	0.9920	0.0038	0.0007	0.9912	0.0071	0.9919	0.0040	0.0007	0.9902	0.0057	0.9918	0.0038	0.0016	0.9911	0.0065	0.9919	0.0038	0.0007
Feb-02	0.9943	0.0084	0.9939	0.0077	-0.0004	0.9943	0.0083	0.9939	0.0077	-0.0003	0.9945	0.0090	0.9940	0.0079	-0.0005	0.9941	0.0087	0.9939	0.0079	-0.0002
Mar-02	0.9965	0.0035	0.9923	0.0058	-0.0041	0.9965	0.0036	0.9925	0.0054	-0.0040	0.9965	0.0041	0.9926	0.0054	-0.0039	0.9962	0.0036	0.9925	0.0053	-0.0037
Apr-02	1.0031	0.0041	1.0022	0.0036	-0.0009	1.0032	0.0044	1.0023	0.0036	-0.0009	1.0030	0.0044	1.0022	0.0035	-0.0008					
May-02	1.0000	0.0057	1.0002	0.0034	0.0002	0.9998	0.0060	1.0001	0.0031	0.0004	0.9996	0.0063	1.0001	0.0031	0.0005					
Jun-02	1.0048	0.0076	1.0035	0.0038	-0.0013	1.0047	0.0076	1.0036	0.0039	-0.0011	1.0047	0.0081	1.0036	0.0040	-0.0011					
Jul-02	0.9992	0.0025	0.9991	0.0026	-0.0001	0.9993	0.0023	0.9991	0.0026	-0.0002	0.9993	0.0022	0.9991	0.0026	-0.0002					
Aug-02	0.9936	0.0074	0.9933	0.0064	-0.0003	0.9935	0.0076	0.9933	0.0065	-0.0002	0.9933	0.0080	0.9933	0.0066	-0.0001					
Sep-02	0.9982	0.0091	0.9979	0.0068	-0.0003	0.9980	0.0094	0.9980	0.0068	0.0000	0.9980	0.0097	0.9980	0.0068	0.0000					
Oct-02	0.9993	0.0035	0.9988	0.0034	-0.0005	0.9993	0.0036	0.9988	0.0034	-0.0005	0.9994	0.0039	0.9988	0.0035	-0.0005					
Nov-02	0.9977	0.0058	0.9981	0.0049	0.0004	0.9976	0.0057	0.9981	0.0049	0.0005	0.9978	0.0060	0.9982	0.0051	0.0004					
Dec-02	0.9968	0.0036	0.9971	0.0024	0.0002	0.9969	0.0036	0.9971	0.0025	0.0002	0.9970	0.0037	0.9971	0.0025	0.0001					
Average		0.0064		0.0055	-0.0002		0.0065		0.0056	-0.0002		0.0066		0.0056	-0.0001		0.0065		0.0060	-0.0003
Average Absolute					0.0008					0.0008					0.0008					0.0009

References

Bailar, B.A. (1989), “Information Needs, Surveys, and Measurement Errors,” in Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (eds.), *Panel Surveys* (pp. 348-374), New York: John Wiley and Sons, Inc.

Binder, D.A. (1998), “Longitudinal Surveys: Why Are These Surveys Different From All other Surveys?,” *Survey Methodology*, **12**, 101-108.

Bureau of Labor Statistics. (2001), Chapter 7, “Estimation,” *Current Employment Statistics Manual*: U.S. Bureau of Labor Statistics, Washington, D.C.

Bureau of Labor Statistics. (2003), “BLS Establishment Estimates Revised to Incorporate March 2002 Benchmarks,” *Employment and Earnings*, U.S. Bureau of Labor Statistics, Washington, D.C.

Bureau of Labor Statistics. (2004a), Chapter 2, “Employment, Hours, and Earnings from the Establishment Survey, *BLS Handbook of Methods*, U.S. Bureau of Labor Statistics, Washington, D.C.

Bureau of Labor Statistics. (2004b), *Technical Notes to Establishment Survey Data Published in Employment and Earnings*, U.S. Bureau of Labor Statistics, Washington, D.C.

Cantwell, P.J., Caldwell, C.V., Hogan, H., and Konschnik, C.A. (1995), “Examining the Revisions in Monthly Trade Surveys Under a Rotating Panel Design,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 567-572.

Carlin, B.P., Polson, N.G, and Stoffer, D.S. (1992), “A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling,” *Journal of the American Statistical Association*, **87**, 493-500.

Cochran, W.G. (1953), *Sampling Techniques*, New York: John Wiley and Sons, Inc.

Copeland, K.R. (2003a), “Nonresponse Adjustment in the Current Employment Statistics Survey,” *Proceedings of the Federal Committee on Statistical Methodology*, (forthcoming).

Copeland, K.R. (2003b), “Reporting Patterns in the Current Employment Statistics Survey,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (forthcoming).

Curtin, R., Presser, S., and Singer, E. (2000), “The Effects of Response Rate Changes on the Index of Consumer Sentiment,” *Public Opinion Quarterly*, 64, 413-428.

Czajka, J. and Hinkins, S. (1993), “Comparing Advance and Final Estimates: 1990 SOI Corporate Sample,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 592-596.

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992), “Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics,” *Journal of Business and Economic Statistics*, 10, 117-131.

David, M.H., Little, R., Samuhel, M., and Triest, R. (1983), “Imputation Models Based on the Propensity to Respond,” *Proceedings of the Section on Business and Economic Statistics, American Statistical Association*, 168-173.

Deming, W.E. and Stephan, F.F. (1940), “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginals Are Known,” *The Annals of Mathematical Statistics*, **11**, 427-444.

Drew, J.H. and Fuller, W.A. (1981), “Nonresponse in Complex Multiphase Surveys,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 623-628.

Duncan, G.J. and Kalton, G. (1978), “Issues of Design and Analysis of Surveys Across time,” *International Statistical Review*, **73**, 97-117.

Eltinge, J.L. (2002), “Diagnostics for the Practical Effects of Nonresponse Adjustment Methods,” in Groves, R.R., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds.), *Survey Nonresponse* (pp. 417-429), New York: John Wiley and Sons.

Gelman, A.R. and Rubin, D.B. (1992), “Inference from Iterative Simulation under Multiple Sequences,” *Statistical Science*, 7, 457-511.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003), *Bayesian Data Analysis*, 2nd ed, New York: Chapman & Hall.

Groves, R.M. (1989), *Survey Errors and Survey Costs*, New York: John Wiley and Sons, Inc.

Groves, R.M. and Couper, M.P. (1998), *Nonresponse in Household Interview Surveys*, New York: John Wiley and Sons, Inc.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), *Sample Survey Methods and Theory*, New York: John Wiley and Sons, Inc.

Harvey, A.C. (1981), *Time Series Models*, Oxford: Phillip Allan Publishers, Ltd.

Hidiroglou, M.A., Sarndal, C.-E., and Binder, D.A. (1995), “Weighting and Estimation in Business Surveys,” in Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (eds.), *Business Survey Methods* (pp. 477-502), New York: John Wiley and Sons, Inc.

Hogan, H., Cantwell, P.J., and Cruz, C. (1997), “Predicting Final Retail Seals Estimates from Advance Reports,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-31.

Judkins, D.R. (1990), “Fay’s Method for Variance Estimation,” *Journal of Official Statistics*, **6**, 223-239.

Kalton, G. (1986), “Handling Wave Nonresponse in Panel Surveys,” *Journal of Official Statistics*, **2**, 303-314.

Kalton, G. and Kasprzyk, D. (1982), “Imputing for Missing Survey Response,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-31.

Kalton, G. and Kasprzyk, D. (1986), “The Treatment of Missing Data,” *Survey Methodology*, **12**, 1-16.

Kalton, G. and Miller, M.E. (1986), “Effects of Adjustments for Wave Nonresponse on Panel Survey Estimates,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 194-199.

Kaspryzyk, D., Duncan, G., Kalton, G., and Singh, M.P. (eds.) (1989), *Panel Surveys*, New York: John Wiley and Sons, Inc.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., and Presser, S. (2000), “Consequences of Reducing Nonresponse in a National Telephone Survey,” *Public Opinion Quarterly*, 64, 125-148.

Kovar, J.G. and Whitridge, P.J. (1995), “Imputation of Business Survey Data,” in Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (eds.), *Business Survey Methods* (pp. 403-423), New York: John Wiley and Sons, Inc.

Lepkowski, J.M. (1989), “Treatment Of Wave Nonresponse In Panel Surveys,” in Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (eds.), *Panel Surveys* (pp. 348-374), New York: John Wiley and Sons, Inc.

Lessler, J.T. and Kalsbeek, W.D. (1992), *Nonsampling Error in Surveys*, New York: John Wiley and Sons, Inc.

Little, R.J.A. (1986), “Survey Nonresponse Adjustments for Estimates of Means,” *International Statistical Review*, **54**, 139-157.

Little, R.J.A. (1993), “Pattern-Mixture Models for Multivariate Incomplete Data,” *Journal of the American Statistical Association*, **88**, 125-134.

Little, R.J.A. and David, M.H. (1983), “Weighting Adjustments for Non-response in Panel Surveys,” Working paper: U.S. Bureau of the Census, Washington, D.C.

Little, R.J.A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data,* 2nd edition, New York: John Wiley and Sons.

Little, R.J.A. and Su, H.-L. (1989), “Item Nonresponse in Panel Surveys,” in Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M.P. (eds.), *Panel Surveys* (pp. 403-423), New York: John Wiley and Sons, Inc.

Madow, L.H. and Madow, W.G. (1978), “On Link Relative Estimators,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 534-539.

Madow, W.G., Nisselson, H., and Olkin, I. (eds.) (1983), *Incomplete Data in Sample Surveys*, New York: Academic Press.

Merkle, D.M., and Edelman, M. (2002), “Nonresponse in Exit Polls: A Comprehensive Analysis,” in Groves, R.R., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds.), *Survey Nonresponse* (pp. 243-257), New York: John Wiley and Sons.

Office of Management and Budget (2001), *Measuring and Reporting Sources of Error in Surveys, Statistical Policy Working Paper 31*, Springfield VA: National Technical Information Service.

Oh, H.L. and Scheuren, F.J. (1983), “Weighting Adjustment For Unit Nonresponse,” in Madow, W.G., Olkin, I., Rubin, D.B. (eds.), *Incomplete Data in Sample Surveys. Vol. 2: Theory and Bibliography* (pp. 143-184), New York: Academic Press, Inc.

Pfeffermann, D. and Nathan, G. (2002), “Imputation for Wave Nonresponse: Existing Methods and a Time Series Approach,” in Groves, R.R., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds.), *Survey Nonresponse*, New York: John Wiley and Sons.

Rao, J.N.K., Srinath, K.P., and Quenneville, B. (1989), “Estimation of Level and Change Using Current Preliminary Data,” in Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (eds.), *Panel Surveys* (pp. 348-374), New York: John Wiley and Sons, Inc.

Rizzo, L., Kalton, G., and Brick, J.M. (1996), “A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse,” *Survey Methodology*, **22**, 43-53.

Rosen, R.J., Clayton, R.L., and Rubino, T.R. (1991), “Controlling Nonresponse in an Establishment Survey,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 587-592.

Rosen, R.J., Clayton, R.L., and Wolf, L.L. (1993), “Long Term Retention of Sample Members under Automated Self-Response Data Collection,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 748-752.

Rosenbaum, P.R. and Rubin, D.B. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects, “ *Biometrika*, 70, 467-474.

Rubin, D.B. (1976), “Inference and Missing Data,” *Biometrika*, **63**, 581-592.

Shao, J. Chen, Y., and Chen, Y. (1998), “Balanced Repeated Replication for Stratified Multistage Survey Data under Imputation,” *Journal of the American Statistical Association,* **93**, 819-831.

Sinharay. S. (2003), “ Assessing Convergence of the Markov Chain Monte Carlo Algorithms: A Review,” Educational Testing Service Research Report.

Solon, G. (1989), “The Value of Panel Data in Economic Research,” in Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (eds.), *Panel Surveys* (pp. 348-374), New York: John Wiley and Sons, Inc.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002), “Bayesian measures of model complexity and fit” (with discussion)*, J. Roy. Statist. Soc. B,* **64**, 583-640.

Werking, G.S. (1997), “Overview of the CES Redesign,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 512-516.

West, S.A. (1983), “A Comparison of Different Ratio and Regression Type Estimators for the Total of a Finite Population,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 388-393.

West, S., Butani, S., Witt, M., and Adkins, C. (1989), “Alternative Imputation Methods for Employment Data,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 227-232.