ESTIMATION OF HOURS AND EARNINGS IN THE CURRENT EMPLOYMENT STATISTICS SURVEY

Stephen M. Woodruff, Bureau of Labor Statistics
441 G St. N.W., Room 2733, Washington D.C. 20212

1. Introduction

The Bureau of labor Statistics' (BLS) Current Employment Statistics (CES) Survey gathers data monthly from over 340,000 nonagricultural business establishments for the purpose of estimating total employment, women and production workers, hours, and earnings. Estimates are made for over 1500 industry cells, complementing the demographic detail provided by estimates of employment from the Current Population Survey. In addition to the monthly estimates of employment level and month to month change in employment which are of primary importance to the users of this data, the estimates of level and change in hours and earnings are becoming increasingly important. This paper examines the current estimators for level and change in the published hours and earnings data. We investigate the data relationships that the current estimators were designed to exploit and suggest some possible minor improvements to the current system.

The sample of business establishments used in this survey is substantially fixed over time and is composed of most large establishments with a less extensive sample of smaller establishments. The CES sample is obtained by soliciting business establishments until a "sufficient" number agree to participate, and thus no sampling distribution (or response mechanism) can be assumed. Variability is largely due to nonresponse, and in the simulation studies to be described later, this nonresponse is the primary source of error in making inferences about hours and earnings for the entire CES population.

The BLS currently uses an estimator called the difference link and taper (LT) for their estimates of Average Weekly Hours (AWH), Average Hourly Earnings (AHE), and Average Weekly Earnings (AWE). These three items are estimated for each of the over 1500 estimating cells in the CES survey and the results published in the Bureaus' monthly "Employment and Earnings" bulletin. In addition, these cell estimates are aggregated to higher level estimates by weighting together the detailed estimates according to their estimated employment or estimated hours (which one depends upon the variable being aggregated). These weights are themselves estimated using the LT.

LT estimation was developed and tested by the BLS at least as early as the 1950s, but very little documentation on the properties of this estimator as it pertains to the CES exists today. One purpose of this paper is the review of this estimator in the current CES environment to insure that it is still appropriate.

A primary source of error in this survey is the result of "data flow", the term used to describe the time dependent arrival of data for processing at the BLS. In short, the firms that report first (and whose data makes up the initial estimates) are often different with respect to the variables being measured from the later reporters and this creates statistical challenges for anyone attempting to make accurate initial estimates from this "unbalanced" early data.

These LT estimates are more difficult to evaluate than the CES survey's estimates for total national employment for which actual population counts become available one to two years after the three closing estimates are produced. The actual population averages that the LT is used to estimate are never known and thus we are forced to evaluate the LT based on actual closing revisions, and theoretical behavior. If the probability models that we use to describe the LT are appropriate then the theoretical evaluation can be very informative. Unfortunately, these theoretical evaluations often miss some important feature of the data that went unnoticed and hence was left out of the model. The greatest possible care must be taken to accurately describe both available data and data relationships. We have no alternative in this case to a theoretical evaluation which will be a combination of deriving expressions for variance and bias (mean square error) and computer simulation by replicating the sampling and estimation process on a known population derived from our CES historical data base.

Section two describes CES data flow, section three discusses current and planned solutions to the estimation problems created by this data flow, section four presents some empirical comparisons of several estimation methodologies, and section five contains conclusions of this research.

2. CES Data Flow

Establishments in the CES sample report employment, hours, and earnings data for the pay period that includes the 12^{th} of each month. They report this data on a shuttle schedule with spaces for six establishment employment variables (all employment (AE), production workers (PW), women workers (WW), total weekly hours for production workers (WH), total weekly payroll (PR), and overtime hours (OT)) over a period of 13 months (December to December). Thus there are 13x6 = 78 entries to complete as the year progresses.

In each of the over 1500 estimating cells the micro data collected on these shuttle schedules is used to estimate AHE and AWH by the LT method. The estimate for AWE is the product of the LT estimates for AHE and AWH. These three averages are estimated at cell level, then further LT estimation is used to estimate the cell weights. A cell's weight is its estimated employment in production workers (PW). These weights are used to aggregate AHE, AWH, and AWE across cells for higher level estimates (2–digit SICs and Industry Divisions).

The first set of estimates for these averages for the current month are preliminary figures based on the initially available microdata that passed editing. These are called the first closing estimates, and are based on 50% to 60% of the sample reports received by the closing date for the current month. Along with these first closing estimates for the current month, second closing estimates for the previous month are computed but with the additional data that has arrived since the first closing estimates for the previous month were computed. These second closing estimates can usually depend on 70% to 80% of the sample reports. Finally, during the current month, third closing estimates for two months in the past can be computed and these are usually based on better than 90% of sample reports. Thus we have three estimates of these averages for any given month, the first closing estimate, the second closing estimate, and the third closing estimate. It is desirable for the differences between these estimates (closing revisions) to be small.

Closing revisions as well as bias and variance in the estimates derived from CES data are effected by two other important features of CES data flow. The first feature is a strong tendency for the same set of units to report for a given closing each month. For example if a given sample unit reported at second closing for June then it will probably report at second closing for July, August and so on. A similar property holds for the other closings. The second feature, related to the first, is that at any time the set of sample units which have reported for month t, have usually also reported for the previous month, t–1. For example, if this is September, then the set of sample units for which August data are available is very nearly contained in the set of sample units for which July data are available.

The effect of this data flow (and sample composition) on bias and variance in the estimates of average hours and earnings is neither fully understood nor documented. This paper will partially answer these problems.

3. The Link and Taper Estimator – Analysis and Alternatives

3.1 Alternatives

The LT is used to estimate three averages or ratios, AHE, AWH, and the production worker ratio (PWR, the ratio of production workers to all employment), the latter is used to weight cell estimates when aggregates across cells are computed. The LT procedure can be described as a way of adjusting the sample ratio estimates for AHE, AWH, or PWR to reduce the effect of peculiarities in the matched responding sample data.

We compare three estimators of the above averages, the first is a simple ratio estimator, and the other two are versions of the difference link and taper (LT) estimator. It may be desirable for the reader to skip most of the detailed derivations and read the main results of this section. These main results are highlighted and follow the four asterisks. The derivation of each result generally precedes it.

Let r_t^k be the k^{th} closing estimate for time t of AHE. Then $r_t^k = [\sum_{i\varepsilon s} p_{it}] / [\sum_{i\varepsilon s} h_{it}] = (p_t/h_t)$, where $p_t = \sum_{i\varepsilon s} p_{it}$, h_t defined similarly, p_{it} is the total payroll in sample establishment i for the week of the 12^{th} of month t, h_{it} is the total hours worked during that week in establishment i, and s is the set of sample establishments with usable data at closing k for employment, payroll, and hours for month t in the particular estimation cell (note that we will not introduce any notation for cell identification because this paper will be restricted to estimation at cell level and the process is identical across cells). The matched sample ratios for PWR and AWH are defined analogously but with different variates in numerator and denominator and so we introduce no extra notation for these. In section four we must distinguish between r_t^k for AWH, and AHE. For the rest of this section assume that r_t^k pertains to AHE (analogous results follow for AWH and PWR). We will consider three estimators for AHE. The first is the straight ratio estimator r_t^k. The second is the LT estimator denoted, LT_t^k and defined as: $LT_t^k = r_t^k + \gamma(LT_{t-1}^{k+1} - r_{t-1}^{k+1})$ where γ=0.9. Here both r_t^k and r_{t-1}^{k+1} are computed from the same set of sample units (i.e. matched across times t and t−1 as well as across the particular data items in numerator and denominator). The third estimator is the LT_t^k with:

$$\gamma = (p_t/p_{t-1})(h_{t-1}/h_t) \qquad 3.1.1$$

the "optimal" value for γ which we derive later in this section.

LT_t^k with γ=0.9 is the version of the link and taper currently used in the CES program.

From now on assume we are making estimates for first closing (k=1) and drop the superscript k, (LT_t^1 becomes LT_t and LT_{t-1}^2 becomes LT_{t-1}) Thus the first closing LT is a function of the first closing r_t and second closing versions of LT_{t-1} and r_{t-1}. The optimal value of γ is determined by minimizing $E(LT_t - \mu_t)^2$ where μ_t is the population mean that we want to estimate. Then:

$$E(LT_t - \mu_t)^2 = E(r_t + \gamma(LT_{t-1} - r_{t-1}) - \mu_t)^2 \qquad 3.1.2$$

These expectations are with respect to probability models that do not include sampling distributions. The sampling distribution is not well documented and the sample is fixed over time. The stochastic structure which defines these expectations and which will be used to derive the optimal value for γ will be defined later in this section.

Historical behavior suggests that the second closing and later LT estimates are fairly stable and so we assume $LT_{t-1}^2 \doteq LT_{t-1}^k$ for all t and k>2. This follows since the second and later closings contain most of the sample data.

Minimizing (3.1.1) with respect to γ and assuming the constancy of LT_{t-1}^2 we get:

$$\gamma_{opt} = \mathrm{Cov}(r_t, r_{t-1}) / [\mathrm{Var}(r_{t-1}) + (LT_{t-1} - \mu_{t-1})^2] \qquad 3.1.3$$

and assuming (or hoping) $LT_{t-1}^2 \doteq \mu_{t-1}$,

$$\gamma_{opt} \doteq \mathrm{Cov}(r_t, r_{t-1}) / \mathrm{Var}(r_{t-1}) \qquad 3.1.4$$

This γ_{opt} can be readily estimated without knowing $\mathrm{Var}(r_{t-1})$ under the model to be discussed next.

Let r_t and r_{t-1} be ratios of total payroll to total hours (average hourly earnings, AHE) at months t and t−1 respectively where the sums in the numerator and denominator of both r_t and r_{t-1} are over the same set, s, of matched sample units (matched across both months t and t−1 and the two data variables, weekly hours and payroll, [WH and PR]). Recall $r_t = p_t/h_t$ where $p_t = \sum_{i\varepsilon s} p_{it}$ and $h_t = \sum_{i\varepsilon s} h_{it}$ and think of $(p_{t-1}, h_{t-1}, p_t, h_t)$ as a vector random variable.

We can adequately capture the stochastic relationships between these random variables in a step wise sequence of dependencies. For notational simplicity, fix the time subscript t and let t−1 be denoted as "p" (past) and let t be denoted as "c" current. The above 4–tuple becomes (p_p, h_p, p_c, h_c) and the we model the relationship between these four components in four stages

1) h_p is assumed to have a mean and variance

2) p_p is substantially explained by h_p according to the following model

$$p_p = \beta_{p_p} h_p + \varepsilon_{p_p} \quad \text{where } E(\varepsilon_{p_p})=0 \text{ and } V(\varepsilon_{p_p})<\infty, \text{ and } \beta_{p_p} \text{ is an unknown constant.}$$

3) The behavior of h_c is largely explained by h_p and the following model:

$$h_c = \beta_{h_c} h_p + \varepsilon_{h_c}, \; E(\varepsilon_{h_c})=0 \text{ and } V(\varepsilon_{h_c})<\infty \text{ and } \beta_{h_c} \text{ is an unknown constant.}$$

4) The behavior of p_c can be explained through p_p and the following model:

$$p_c = \beta_{p_c} p_p + \varepsilon_{p_c}, \text{ where } E(\varepsilon_{p_c})=0 \text{ and } V(\varepsilon_{p_c})<\infty, \; \beta_{p_c} \text{ is an unknown constant.}$$

Finally assume that h_p and all these εs are stochastically independent.

Given this model we have:

$$\mathrm{Cov}(r_p, r_c) =$$

$$\mathrm{Cov}([\beta_{p_p} h_p + \varepsilon_{p_p}]/h_p, \; [\beta_{p_c} p_p + \varepsilon_{p_c}]/[\beta_{h_c} h_p + \varepsilon_{h_c}])$$

$$= \mathrm{Cov}(\beta_{p_p} + \varepsilon_{p_p}/h_p, \; [\beta_{p_c}\{\beta_{p_p} h_p + \varepsilon_{p_p}\} + \varepsilon_{p_c}]/[\beta_{h_c} h_p + \varepsilon_{h_c}])$$

$=Cov(\varepsilon_{p_p}/h_p,[\beta_{p_c}\beta_{p_p}h_p+\beta_{p_c}\varepsilon_{p_p}+\varepsilon_{p_c}]/[\beta_{h_c}h_p+\varepsilon_{h_c}])$

$=Cov(\delta_h, [\beta_{p_c}\beta_{p_p} + \beta_{p_c}\delta_h + \delta_w]/[\beta_{h_c} + \delta_t])$

where $\delta_h = \varepsilon_{p_p}/h_p$, $\delta_t = \varepsilon_{h_c}/h_p$, and $\delta_w = \varepsilon_{p_c}/h_p$.

By the multi linearity of the Cov(.) function we have:

$Cov(r_p,r_c) = [\beta_{p_c}\beta_{p_p}/\beta_{h_c}]Cov(\delta_h, 1/[1 + \delta_t/\beta_{h_c}) +$

$[\beta_{p_c}/\beta_{h_c}]Cov(\delta_h, \delta_h[1/\{1 + \delta_t/\beta_{h_c}\}]) +$

$[1/\beta_{h_c}]Cov(\delta_h, \delta_w/[1 + \delta_t/\beta_{h_c}])$

Now, applying a geometric expansion to the second term in each covariance function we get a linear approximation to this covariance as:

$Cov(r_p,r_c) = [\beta_{p_c}\beta_{p_p}/\beta_{h_c}]Cov(\delta_h, 1 - \delta_t/\beta_{h_c}) +$

$[\beta_{p_c}/\beta_{h_c}]Cov(\delta_h,\delta_h-\delta_h\delta_t/\beta_{h_c})+$

$[1/\beta_{h_c}]Cov(\delta_h,\delta_w-\delta_w\delta_t/\beta_{h_c})$

$= -[\beta_{p_c}\beta_{p_p}/\beta^2_{h_c}]Cov(\delta_h,\delta_t) + [\beta_{p_c}/\beta_{h_c}]\sigma^2_{\delta_h} -$

$[\beta_{p_c}/\beta^2_{h_c}]Cov(\delta_h, \delta_h\delta_t) + [1/\beta_{h_c}]Cov(\delta_h,\delta_w) -$

$[1/\beta^2_{h_c}]Cov(\delta_h, \delta_w\delta_t).$

By conditioning on h_p, and computing the four covariances in this last equality by the conditional covariance formula, we find that the four covariances are zero.

For example: $Cov(\delta_h,\delta_t) = Cov(E(\delta_h|h_p),E(\delta_t|h_p)) + E(Cov(\delta_h,\delta_t|h_p))$

$=Cov(E(\varepsilon_{p_p}/h_p|h_p),E(\varepsilon_{h_c}/h_p|h_p)) +$

$E(Cov(\varepsilon_{p_p}/h_p,\varepsilon_{h_c}/h_p|h_p))$

$=Cov(0,0) + E(0)$ by independence of the εs and h_p

$=0.$

Therefore:

$Cov(r_p,r_c) = [\beta_{p_c}/\beta_{h_c}]\sigma^2_{\delta_h} = (\beta_{p_c}/\beta_{h_c})V(\varepsilon_{p_p}/h_p) = (\beta_{p_c}/\beta_{h_c})V(p_p/h_p) = (\beta_{p_c}/\beta_{h_c})V(r_p).$

Thus $\gamma_{opt} = Cov(r_p,r_c)/V(r_p)$, is $(\beta_{p_c}/\beta_{h_c})$.

****** 1)** If payroll and weekly hours move at about the same rates in the same direction then $\beta_{p_c} \doteq \beta_{h_c}$ and the optimal coefficient, $\gamma_{opt} = \beta_{p_c}/\beta_{h_c}$, is close to unity. However, a more direct estimate of this ratio is $\hat{\beta}_{p_c}/\hat{\beta}_{h_c}$, where $\hat{\beta}_{p_c} = p_c/p_p$ and $\hat{\beta}_{h_c} = h_c/h_p$.

Letting $\gamma = \hat{\gamma}_{opt} = \hat{\beta}_{p_c}/\hat{\beta}_{h_c}$ 3.1.5

is probably a better way to let the observed data have a more direct influence on the estimator than simply fixing this ratio at unity or 0.9.

LT_c with $\gamma = \hat{\beta}_{p_c}/\hat{\beta}_{h_c}$ is the third estimator to be evaluated in this paper.

****** 2)** This estimator reduces to the link relative estimator given by 3.1.6.

$LT_c = LT_p(r_c/r_p).$ 3.1.6

This is the estimator currently used for first closings in the CES program.

First closing is where the maximum potential benefit of link and taper is achieved over it's simple alternative, r_c. Thus current applications of link and taper in the CES program are probably appropriate.

3.2 Properties

Both LT_c can be thought of as a modification of the regression estimator under bivariate normality. Suppose that

$\begin{bmatrix} r_c \\ r_p \end{bmatrix} \sim N\left[\begin{bmatrix} \mu_c \\ \mu_p \end{bmatrix}, \begin{bmatrix} \sigma^2_c & C_{c,p} \\ C_{p,c} & \sigma^2_p \end{bmatrix}\right]$, then:

$(r_c|r_p) \sim$

$N(\mu_c-(C_{p,c}/\sigma^2_p)(\mu_p-r_p),\sigma^2_c-(C^2_{p,c}/\sigma^2_p)).$

In order to condition on r_p, we need μ_p and the covariance matrix. Then we have:

$E(r_c + (C_{p,c}/\sigma^2_p)(\mu_p - r_p) | r_p) = \mu_c$ and

$V(r_c + (C_{p,c}/\sigma^2_p)(\mu_p - r_p) | r_p)$

$=\sigma^2_c-(C^2_{p,c}/\sigma^2_p).$ 3.2.1

Thus $R_c = r_c + (C_{p,c}/\sigma^2_p)(\mu_p - r_p)$, the regression adjusted estimator, has the desired (μ_c) expected value (conditional on r_p) and a conditional variance of $\sigma^2_c - C^2_{p,c}/\sigma^2_p$. It is appropriate to condition on r_p since we are assuming that $LT_p \doteq \mu_p$.

****** 3)** To the extent that LT_p is close to μ_p and γ is close to $(C_{p,c}/\sigma^2_p)$, the link and taper procedures will have smaller variance and bias than r_c. Heuristically, we might say that the link and taper procedures capture mathematically the tendency for r_c to be consistently related to μ_c over time the way r_p is related to μ_p, or considering CES data flow and response imbalance due to closing, if r_c consistently underestimates μ_c, the link and taper procedures will adjust for such tendencies to the extent that μ_p and $(C_{p,c}/\sigma^2_p)$ can be accurately estimated.

We now consider the effect on mean square error of replacing $C_{p,c}/\sigma^2_p$ and μ_p in the regression adjusted estimator with estimates of these quantities.

$LT_c = r_c + \gamma(LT_p - r_p).$

Adding and subtracting $C_{p,c}/\sigma^2_p$ and μ_p the LT_c becomes:

$LT_c=$

$r_c+(\gamma+C_{p,c}/\sigma^2_p-C_{p,c}/\sigma^2_p)(LT_p+\mu_p-\mu_p-r_p)$

$= r_c + (C_{c,p}/\sigma^2_p)(\mu_p - r_p) + \gamma(LT_p - r_p)$

$+ (C_{p,c}/\sigma^2_p)(r_p - \mu_p)$

$= R_c + \gamma(LT_p - r_p) + (C_{p,c}/\sigma_p^2)(r_p - \mu_p)$ 3.2.2

where:

$R_c = r_c + (C_{p,c}/\sigma_p^2)(\mu_p - r_p)$, the regression adjusted estimator when the mean (μ_p) and covariance matrix (or $C_{p,c}$ and σ_p^2) are known. Next note that the term:

$$\gamma(LT_p - r_p) = \gamma^n(LT_{c-n} - r_{c-n})$$

where c – n denotes n months prior to the current month and for large n (in the CES, $14 \le n \le 27$) and $\gamma = 0.9 < 1.0$, this term becomes small and we have:

$$LT_c \doteq R_c + (C_{p,c}/\sigma_p^2)(r_p - \mu_p) \quad \text{for } \gamma < 1.$$

By second closing enough data has arrived so that the second closing link and taper estimates are relatively close to μ_p, making it more appropriate to condition on the estimates, r_p, for the previous time period (or earlier). Thus, conditional mean square error of the LT_c given r_p is a justifiable measure of error and:

$MSE(LT_c|r_p) = V(LT_c|r_p) + Bias^2(LT_c|r_p)$ and conditional on r_p, $(C_{p,c}/\sigma_p^2)(r_p - \mu_p)$ is constant (the bias). Therefore:

$$MSE(LT_c|r_p) = V(R_c|r_p) + (C_{p,c}^2/\sigma_p^4)(r_p - \mu_p)^2.$$

From 3.2.3 we have:

$$MSE(LT_c|r_p) \doteq \sigma_c^2 - (C_{p,c}^2/\sigma_p^2) + (C_{p,c}^2/\sigma_p^4)(r_p - \mu_p)^2 \qquad 3.2.3$$

and from 3.2.1:

$MSE(r_c|r_p) = V(r_c|r_p) + Bias^2(r_c|r_p)$ where:

$V(r_c|r_p) = \sigma_c^2 - (C_{p,c}^2/\sigma_p^2)$, and $Bias(r_c|r_p) = (C_{p,c}/\sigma_p^2)(r_p - \mu_p)$.

Therefore:

$$MSE(r_c|r_p) = \sigma_c^2 - (C_{p,c}^2/\sigma_p^2) + (C_{p,c}^2/\sigma_p^4)(r_p - \mu_p)^2 \qquad 3.2.4$$

****** 4) In summary, 3.2.3 and 3.2.4, say that conditional on r_p, the mean square errors of LT_c and r_c are, for practical purposes, the same. This also follows somewhat heuristically from the expression:**

$(LT_c - r_c) = \gamma^n(LT_{c-n} - r_{c-n})$ which goes to 0 as n gets large for $\gamma < 1$.

Note that when we choose $\gamma > 1$ the term, $\gamma(LT_p - r_p)$, is no longer necessarily negligible. In this case, from 3.2.4, the mean square error of LT_c becomes:

$$MSE(LT_c|r_p) \doteq \sigma_c^2 - (C_{p,c}^2/\sigma_p^2) + (C_{c,p}^2/\sigma_p^4)(r_p - \mu_p)^2 + \gamma^2(LT_p - r_p)^2 + 2\gamma(C_{p,c}/\sigma_p^2)(LT_p - r_p)(r_p - \mu_p) = MSE(LT_c|r_p)_{\gamma=.9} + Q \text{ where:}$$

$Q = \gamma^2(LT_p - r_p)^2 + 2\gamma(C_{p,c}/\sigma_p^2)(LT_p - r_p)(r_p - \mu_p)$ and $MSE(LT_p|r_p)_{\gamma=.9}$ is the mean square error of the link and taper when $\gamma = .9$ (and this $MSE \doteq MSE(r_c|r_p)$).

Here Q is minimized when $\gamma = (C_{p,c}/\sigma_p^2)([r_p - \mu_p]/[r_p - LT_p])$ 3.2.5

For this value of γ we have:

$Q = -(C_{p,c}^2/\sigma_p^4)(r_p - \mu_p)^2 \le 0$. This last expression shows that when $r_p = \mu_p$, then r_c has the smallest MSE.

Note that this optimal value for γ is different from 3.1.4. This is due to minimizing the respective squared differences with respect to conditional (3.2.5) versus unconditional (3.1.4) distributions. When $LT_p = \mu_p$, then both these expressions for the optimal γ are the same. In several years, the Bureau will have benchmark data needed to compute the population parameter, μ_p, and verify the assumption $LT_p = \mu_p$. Note that if this assumption is false, then this planned benchmark data can be used to estimate both the conditional and unconditional versions of the optimal γ.

For the LT_c to be an improvement over r_c it suffices that $Q < 0$. Since $\gamma = \hat{\beta}_{p_c}/\hat{\beta}_{h_c} \doteq (C_{p,c}/\sigma_p^2)$ for the last two versions of the link and taper, $Q < 0$ occurs when:

$$(LT_p - r_p)^2 + 2(LT_p - r_p)(r_p - \mu_p) < 0.$$

But this expression can be rewritten:

$$(LT_p - \mu_p)^2 - (r_p - \mu_p)^2 < 0.$$

This holds whenever LT_p is closer to μ_p than is r_p. A perfectly reasonable requirement!

****** 5) That is, if the link and taper with $\gamma = 1.0$ or $\hat{\beta}_{p_c}/\hat{\beta}_{h_c}$, did a better job of estimating last month's average then it will do better (on average) this month for estimating μ_c. This also means that when LT_p is farther from μ_p than is r_p, then $MSE(LT_c|r_p)_{\gamma=1} > MSE(r_c|r_p)$, (it does worse, on average, for estimating μ_c).**

4. A Simulation Study

This simulation verifies the theoretical results derived in the previous section and exposes some possible problems with using an unstratified combined ratio estimator (r_t) for Average Weekly Hours.

Two test universes were constructed from 1985 CES sample data in SICs 5211 and 5231. A sample was selected and fixed for each replication of the simulated CES data flow and the estimators were computed for first closing over a four month period. One simulation run was done for second closing to see if there were any important differences between first and second closing behavior. Except for expected reductions in MSE, the phenomena observed at first closing was still apparent at second closing. Mean Square Error was estimated by averaging (over the replications of the CES data flow) the squared differences between each estimator and the actual population value being estimated. MSE is tabulated for each estimator for the four months, May, June, July, and August (M,J,J,A). Bias is tabulated when it is a significant component of MSE and highlighted directly under MSE.

The three estimators for which MSE (and bias) were estimated are:

1) r_t, the plain ratio given in the third paragraph of section three,

2) $LT_{.9}$ which is LT_t (or LT_c) with $\gamma = .9$,

and 3) LT_o, which is LT_c with $\gamma = \gamma_{opt}$ as given in 3.1.1, 3.1.4, or 3.1.5.

Consider Table 1. For the month of June, r_t has an estimated MSE of .109 for Average Hourly Earnings and this MSE is mostly variance (since there is no bias highlighted

directly beneath this figure). For Average Weekly Hours, r_t has an estimated MSE of 2.37 in June and for this variable r_t has a negative bias, −1.35.

A stratified sample was selected using roughly the same unit selection probabilities given in the CES State Operating Manual for SIC 52. They are as follows:

Jan All Employment	Selection Probability
250 +	1
[100,249]	0.6
[50,99]	0.4
[20,49]	0.1
[10,19]	0.05
[0,9]	0.018

The sample indicator functions, $\{I_i\}_{i=1}^{N}$, are independent, $\Pi_{ij}=\Pi_i\Pi_j$ (where $E(I_i)=\Pi_i$ and $E(I_iI_j)=\Pi_{ij}$), and $n = \sum_{i=1}^{N} I_i$, the sample size, is a random variable.

Heirarchal data flow was simulated using Markov transition probability matrices. The $(i,j)^{th}$ element in these matrices is the probability of an establishment having a closing code of j for an arbitrary month (a) given it had a closing code of i for the previous month (a−1). Such matrices have stationery distributions which also are the unconditional probability distributions of the closing codes. These stationary distributions vary with the establishment size and thus an establishment's closing transition matrix is a function of its size (Mar 1985 All Employment). The closing transition matrices are as follows:

Mar All Employment		Closing Transition Matrix
1.	[0,49]	$\begin{bmatrix} .75 & .19 & .05 & .01 \\ .42 & .42 & .11 & .05 \\ .43 & .39 & .14 & .04 \\ .18 & .41 & .18 & .23 \end{bmatrix}$
2.	[50,249]	$\begin{bmatrix} .75 & .19 & .05 & .01 \\ .42 & .42 & .11 & .05 \\ .13 & .59 & .21 & .07 \\ 0.0 & .15 & .37 & .48 \end{bmatrix}$
3.	250 +	$\begin{bmatrix} .75 & .19 & .05 & .01 \\ .22 & .52 & .22 & .04 \\ .23 & .52 & .19 & .06 \\ .06 & .25 & .30 & .39 \end{bmatrix}$

For size group 1, the unconditional or long term probability of units in this size group having a closing code of one for an arbitrary month is .62, for group two it is .55, and for group three it is .46. This roughly follows historical experience. For an establishment in group two (All Employment between 50 and 249) the probability of having a closing code of 3 this month given it was 1 last month is the (1,3)–element of the transition matrix, .05.

Note that for totally heirarchal data flow the (3,1), (4,1), and (4,2) elements must be zero. For the above matrices they are not zero but they generate heirarchal data flow 95% to 98% of the time, as observed historically.

The universe size for SIC 5211 was N=1063 and the sample size was n=138 for Tables one, two, and three. As a verification that the phenomena observed in these three tables was not due to an extreme sample, a second sample was selected independently of the first using these selection probabilities and the simulation study was rerun with this new sample of size n=121. These results are contained in Table 4 and are consistent (allowing for different sample size) with those observed on the sample of size 138 (Tables one, two, and three). Each table in this section contains simulation results based on 100 replications of CES data flow.

Table 2 is an independent replicate of Table 1. That is, it is derived from 100 additional independent replicates of CES data flow and the same sample of 138 units. When these two tables are compared, we get a rough estimate of the standard error of the entries in these tables. It would be appropriate to average table one and two to obtain a third estimate of simulation MSE and bias with 1/2 the variance of the entries in either table one or two. This rough estimate of standard error indicates no very statistically significant difference between the three estimators tested. If more than 100 replicates are required to demonstrate statistical significance, then any difference in performance is probably to small to be of practical significance (i.e. this is a large survey and system changes are very costly).

Table five summarizes simulation results for SIC 5231 where the universe size is 145 and the sample size is 12, otherwise the simulation is analogous to table one. Note the sign change on the bias of Average Weekly Hours (AWH) for this SIC versus SIC 5211. Except for their magnitudes, the trends observed in tables one through four are similar to those observed in table five.

The simulation study is consistent with the theoretical results derived in the previous section.

r_t is computed throughout the simulation as a combined ratio estimator (without regard to stratification). This presents no problem for AHE but for AWH there is a severe bias present in every case tested. Apparently small establishments behave differently from large ones with respect to AWH. When estimation is done by size strata, this bias will probably become inconsequential. This is partially demonstrated in table 6 where the estimators were compared in SIC 5211 for the sample units with March AE < 100. The bias, as a percent of MSE, in tables 1 through 4 for AWH is less than in table 6 (this is usually over 80% in tables one to four and about 50% in table 6).

This study is indicative but not definitive for the following reasons. First, estimation is carried on from fourteen to twenty seven months past the benchmark month (rather than one to five). Second, actual closing codes are not used. Based on what we know about CES data flow we hope that we can realistically mimic it in the simulation study. Third, this study was conducted with CES historical data from an extremely small set of industries. Fourth, the CES sample which is our simulation universe is much less skewed (the All Employment variable) than the actual CES universe. Fifth, these results will be conditional on the sample selected.

For these reasons it may be necessary to continue such studies beyond what is covered in this short document.

5. CONCLUSIONS

No substantial differences between the three estimators (r_t, $LT_{.9}$, and LT_0) were found in the simulation study or in the analysis done in section three. The current version of the Bureau's link and taper estimator ($LT_{.9}$) seemed to do slightly better for AHE than the others, but this improvement was so small that it was not statistically significant (at α=.05 level) with the 100 replications of CES data flow used in the simulation study. It may be that the difference link and taper technique can handle outliers better than r_t and this was not tested in either section three or four. Based on the minimal amount of testing done in this study, the difference link and taper (LT_0, $LT_{.9}$) has minimal effect on error reduction. Thus, there is no compelling reason to change the current system.

All three estimators were tested without regard to size stratification This presents no problem for AHE, for AWH there is a severe bias present in every case tested. Apparently small establishments behave differently from large ones with respect to AWH and in such cases estimation must be done by size strata within industry strata to minimize these biases.

REFERENCES

Madow, William G. and Madow, Lillian H.,(1979) "On Link Relative Estimators II", Proceedings of the Section on Survey Research Methods of the American Statistical Association, 1979.

Royall, Richard M. (1981), "The Role of Probability Models in 790 Survey Design and Estimation" BLS Contract Report 80–98.

Table 1. First Closing MSEs For SIC 5211

Average Hourly Earnings

Month Est	M	J	J	A
r_t	.103	.109	.098	.099
$LT_{.9}$	.104	.088	.081	.080
LT_o	.104	.095	.097	.105

Average Weekly Hours

[MSE/(Bias)]

r_t	3.97	2.37	2.02	3.30
	−1.86	−1.35	−1.20	−1.54
$LT_{.9}$	3.94	2.44	2.16	3.14
	−1.86	−1.39	−1.29	−1.55
LT_o	3.94	2.49	2.30	3.37
	−1.86	−1.40	−1.30	−1.58

Table 2. First Closing MSEs For SIC 5211

(Replicate of Table 1.)

Average Hourly Earnings

Month Est	M	J	J	A
r_t	.102	.076	.070	.091
$LT_{.9}$	.115	.103	.092	.079
LT_o	.115	.117	.122	.120

Average Weekly Hours

[MSE/(Bias)]

r_t	3.89	2.53	2.23	3.51
	−1.80	−1.38	−1.29	−1.61
$LT_{.9}$	3.73	2.36	2.02	3.03
	−1.72	−1.30	−1.19	−1.50
LT_o	3.73	2.39	2.11	3.15
	−1.72	−1.29	−1.17	−1.48

Table 5. First Closing MSEs For SIC 5231

Average Hourly Earnings

[MSE/(Bias)]

Month Est	M	J	J	A
r_t	.311	.241	.294	.262
	−.32	−.26	−.34	−.30
$LT_{.9}$	.320	.269	.270	.215
	−.31	−.23	−.26	−.21
LT_o	.320	.288	.316	.288
	−.31	−.23	−.26	−.19

Average Weekly Hours

[MSE/(Bias)]

r_t	4.89	6.07	4.80	2.93
	1.80	2.01	1.63	0.96
$LT_{.9}$	5.05	5.87	5.98	4.11
	1.79	1.99	1.56	0.71
LT_o	5.05	6.19	6.67	4.98
	1.79	2.02	1.59	0.74

Table 3. Second Closing MSEs For SIC 5211

Average Hourly Earnings

Month Est	M	J	J	A
r_t	.023	.022	.025	.020
$LT_{.9}$	.024	.020	.018	.017
LT_o	.024	.023	.024	.026

Average Weekly Hours

[MSE/(Bias)]

r_t	4.49	2.52	2.01	3.47
	−2.08	−1.55	−1.37	−1.79
$LT_{.9}$	4.49	2.70	2.31	3.66
	−2.07	−1.59	−1.47	−1.86
LT_o	4.49	2.75	2.42	3.82
	−2.07	−1.60	−1.49	−1.88

Table 4. First Closing MSEs For SIC 5211

(Second Sample)

Average Hourly Earnings

Month Est	M	J	J	A
r_t	.130	.100	.101	.113
$LT_{.9}$	.126	.103	.089	.082
LT_o	.126	.114	.111	.119

Average Weekly Hours

[MSE/(Bias)]

r_t	7.89	5.70	5.33	7.39
	−2.70	−2.24	−2.11	−2.55
$LT_{.9}$	7.82	5.30	4.84	6.84
	−2.69	−2.18	−2.07	−2.47
LT_o	7.82	5.31	4.91	6.95
	−2.69	−2.17	−2.07	−2.47

Table 6. First Closing MSEs For SIC 5211 and the <100s

Average Hourly Earnings

[MSE/(Bias)]

Month Est	M	J	J	A
r_t	.081	.064	.082	.060
	.199	.171	.209	.158
$LT_{.9}$	.087	.058	.071	.047
	.201	.138	.193	.141
LT_o	.087	.063	.078	.060
	.201	.135	.187	.134

Average Weekly Hours

[MSE/(Bias)]

r_t	.778	.335	.365	.304
	−.69	−.17	.205	−.02
$LT_{.9}$	.830	.420	.322	.347
	−.72	−.25	.16	−.10
LT_o	.831	.472	.421	.503
	−.72	−.26	.140	−.12