A COMPARISON OF SOME ESTIMATORS IN SAMPLING FOR A TIME SERIES WITH LINEAR TREND

Stephen M. Woodruff, Bureau of Labor Statistics

I. INTRODUCTION

This paper deals with the problem of estimating population totals for a population that is changing with time. The problem is to estimate these totals at each time k, where this set of times may correspond to weeks, months or years. These population totals can be thought of as the realization of a time series.

This problem of estimating the realization of a time series using only a sample of the population which is itself fixed over time has been discussed by Madow and Madow (1) and Royall (2). The techniques suggested in these papers assume the actual population total at time k=1 is given. Then the sample data at subsequent times are used to make an estimate of change for the entire population. This estimate of change is then used to derive an estimate of the current population total from an estimate of a previous population total at the preceeding time. The sample that is used to estimate this change is selected at time k=1 and, except for nonresponse, is the sample which is used at all subsequent times.

The purpose of this paper is to investigate a class of estimators and compare their mean square errors. This class of estimators consists of weighted regression estimators where the weight is a function of a single real number, d, in the interval (-1, 2). The other determining variable, h, is the amount of historical data used in the estimator. Thus for each pair (d,h), where $d \in (-1, 2)$ and h is a positive integer, there is an estimator which will be defined later.

The purpose of this paper is to demonstrate, for the particular sampling situation and model assumptions described in this paper, that there is one estimator among the class described above that is best with respect to both mean square error and robustness.

II. DESCRIPTION OF THE POPULATION AND THE ESTIMATORS

The population is assumed to be changing according to the linear model described in this section.

The following notation will be used to describe the stochastic structure.

1) Let $Y_k(i)$ be the random variable associated with the i^{TH} population unit at time k.
2) Let $y_k(i)$ be the realization of $Y_k(i)$.

For each time period k we wish to estimate:

$$\sum_{i \in S} y_k(i)$$

where the summatation is over the entire population denoted by S. The model states that, conditional upon the realizations up to and including time k-1, the expected value of $Y_k(i)$ is proportional to $y_{k-1}(i)$ and that the stochastic processes $Y_k(i)$, $Y_k(j)$, k=1,2,--- are conditionally uncorrelated when $i \neq j$. These statements are expressed algebraically in (2.1) below.

$$E(Y_k(i)/k-1) = \beta_k y_{k-1}(i) \qquad (2.1)$$
$$\text{Cov}(Y_k(i), Y_k(j)/k-1) = \begin{cases} V_k(i) & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$$

The notation for conditional expectation used throughout this paper is $E(\cdot/k-1)$. Estimators based on a variety of models for $V_k(i)$ will be considered later in this section.

Let s_k be the sample at time k and let r_k be it's complement in S. Thus $S = s_k \cup r_k$. Note that in spite of the assumption of a fixed sample over time subscripts appear on s_k and r_k. Because of nonresponse in the sample there is variation within the fixed sample, s, from time period to time period.

The estimator for time period k which is a function of the samples s_k, s_{k-1} --- s_2, $s_1=S$ is denoted:

$$t_k(S) = \sum_{i \in S} t_k(i)$$

where $t_k(i)$ is the estimator for $y_k(i)$.

The estimators that I consider here are primarily of the type:

$$t_k(S) = t_k(s_k) + t_k(r_k)$$

where $t_k(s_k) = \sum_{i \in s_k} y_k(i) = y_k(s_k)$

since $t_k(i) = y_k(i)$ for $i \in s_k$

and $t_k(r_k) = \sum_{i \in r_k} t_k(i)$

That is, $t_k(S)$ is the sum of the observed sample sum at time k and the sum of estimates t of nonsample y values at time k. The estimation of this latter quantity can be classified according to the variance model and this is discussed next.

An estimate for $y_k(r_k)$ can be sought by first estimating β_k (the index of change between time k-1 and time k) and then multiplying this estimate by $t_{k-1}(r_k)$. One way to estimate β_k, suggested by ordinary least square regression, would be to let its estimator be that value of c which minimizes:

$$\sum_{i \in s_k s_{k-1}} (y_k(i) - c\, y_{k-1}(i))^2 \qquad (2.2)$$

where $s_k s_{k-1} = s_k \cap s_{k-1}$. This notation for set intersection will be used for the rest of this paper.

When $V(Y_k(i)/k-1)$ is independent of $y_{k-1}(i)$ then this procedure is reasonable. If however, this is not the case then some account of this dependence on $y_{k-1}(i)$ needs to be reflected in the above sum. For example, if $V(Y_k(i)/k-1)$ is proportional to $y_{k-1}(i)$ then the terms in the above sum (2.2) with smaller $y_{k-1}(i)$ should be weighted more heavily since the pairs $(y_{k-1}(i), y_k(i))$, in such terms, would be expected to lie nearer the true regression line than terms with large $y_{k-1}(i)$.

This weighting can be effectively achieved by transforming the regression variables so that variances are constant. Instead of finding the line through the origin (0,0) which is closest to the set of point's $\{(y_{k-1}(i), y_k(i)), i \in s_k s_{k-1}\}$, find that line which is closest to $\{(y_{k-1}(i)/\sqrt{V(Y_k(i)/k-1)}, y_k(i)/\sqrt{V(Y_k(i)/k-1)}), i \in s_k s_{k-1}\}$.

Note that $V((Y_k(i)/\sqrt{V(Y_k(i)/k-1)})/k-1) = 1$ for all i (thus it is independent of $y_{k-1}(i)$). With this latter set of points the sum to minimize becomes:

$$\sum_{i \in s_k s_{k-1}} (1/V(Y_k(i)/k-1))\,(y_k(i) - c\, y_{k-1}(i))^2$$

In this sum each term is weighted inversely to its variance as desired. The value of c which minimizes this is:

$$c = \frac{\sum_{s_k s_{k-1}} (1/V(Y_k(i)/k-1))\, y_k(i)\, y_{k-1}(i)}{\sum_{s_k s_{k-1}} (1/V(Y_k(i)/k-1))\, y^2_{k-1}(i)}$$

Now if $V(Y_k(i)/k-1) = K_k y^d_{k-1}(i)$ for some constant, K_k, and for all i then

$$c = \left(\sum_{s_k s_{k-1}} y_k(i)\, y^{1-d}_{k-1}(i)\right) / \left(\sum_{s_k s_{k-1}} y^{2-d}_{k-1}(i)\right) \qquad (2.3)$$

Therefore if the conditional variance is proportional to some power of the outcome for the preceeding time period then this c in (2.3) is an unbiased estimator for β_k.

The problem of estimating products of consecutive βs, is much more difficult in general. This is the problem of estimating change between time k-f and time k for f>1. If $V(Y_k(i)/k-1)$ is proporitonal to $y_{k-1}(i)$ (as above with d=1) then an easy solution is available. This is because, $V(Y_k(i)/k-1)$ proportional to $y_{k-1}(i)$ implies $V(Y_k(i)/k-f)$ proportional to $y_{k-f}(i)$. In this case an estimate of change is:

$$c = \left(\sum_{s_k s_{k-f}} y_k(i) \Big/ \sum_{s_k s_{k-f}} y_{k-f}(i)\right)$$

Unfortunately for $d \neq 1$ it does not follow that $V(Y_k(i)/k-1)$ proportional to $y^d_{k-1}(i)$ implies $V(Y_k(i)/k-f)$ is proportional to $y^d_{k-f}(i)$. In this case the expression for c becomes difficult to estimate and therefore when $d \neq 1$ estimators using more than the two time periods k and k-1 will not be considered in this paper.

III. DEFINITION OF ESTIMATORS

Let $$a(p,q,d) = \sum_{i \epsilon s_p s_q} y^{1-d}_p(i)\, y^{1-d}_q(i) \Big/ \sum_{i \epsilon s_p s_q} y^{2-d}_q(i)$$

In situations where confusion is possible $t_k(S)$ will be prefixed by an abbreviated name and semicolon in order to distinguish between different estimators. With this notation the estimators that are considered can be defined as follows:

1) An estimator, R1, based on unweighted regression through the origin (d=0) was tested and is defined as follows:

   R1: $t_k(S) = y_k(s_k) + a(k,k-1,0) \cdot t_{k-1}(r_k)$

2) The link relative estimator (1) was also tested. It is the one currently being used by the Bureau of Labor Statistics for estimation of population totals for the sampling scheme outlined in the second paragraph of the Introduction.

   E0: $t_k(S) = a(k,k-1,1) \cdot t_{k-1}(S)$

3) The next set of estimators E1 through E3 differ only in the amounts of past time period data that are used. They correspond to a weighted regression through the origin with d=1.

   E1: $t_k(S) = y_k(s_k) + a(k,k-1,1)\, t_{k-1}(r_k)$

   E2: $t_k(S) =$ E1: $t_k(S)$ for $k = 2$

   E2: $t_k(S) = y_k(s_k) + a(k,k-1,1) t_{k-1}(s_{k-1} r_k) + a(k,k-2,1)\, t_{k-2}(r_k r_{k-1})$ for $k > 2$

   E3: $t_k(S) =$ E2: $t_k(S)$ for $k=2$ & 3

   E3: $t_k(S) = y_k(s_k) + a(k,k-1,1) t_{k-1}(r_k s_{k-1}) + a(k,k-2,1)\, t_{k-2}(r_k r_{k-1} s_{k-2}) + a(k,k-1,1) t_{k-1}(r_k r_{k-1} r_{k-2})$ for $k > 3$

   This set of estimators (E1 -- E3) can be summarized by saying that if possible a unit's estimate for time k is an update of a known y value from the last time it responded but if it hadn't been a respondent in the near past then it's estimate at time k is an update from it's estimate at time k-1.

4) The next estimator to be considered is the only "true" regression type estimator. All of the above estimators apply a regression coefficient (update factor a ($\cdot$, $\cdot$, $\cdot$)) to an estimate of the auxillary variable. If an estmator for time k is derived by updating from the benchmark month (k=1) then the auxillary variable is known exactly.

   R2: $t_k(S) = y_k(s_k) + a(k,1,1) \cdot y_1(r_k)$
   from the benchmark month (k=1) then the auxillary variable is known exactly.

   R2: $t_k(S) = y_k(s_k) + a(k,1,1) \cdot y_1(r_k)$

5) The Horwitz-Thompson estimator was included in this study as an additional basis for comparing the probability sampling approach with super population models. It is described in detail in part V "Description of Simulation." It is denoted as HT.

6) The final estimator that is considered here is a composite of the Horwitz-Thompson estimator (HT) and E2. The problem of defining and measuring variances in a meaningful way for both HT and E2 has yet to be solved. Since the optimal weights of a composite estimator are a function of these variances another weighting scheme was tried. When there are relatively severe deviations from the superpopulation model, $E(Y_k(i)/k-1) = \beta_k y_{k-1}(i)$, then the regression type estimators deteriorate quickly as time passes until their mean square errors are larger than the mean square error of the Horwitz-Thompson estimator.
   This observation suggested the following estimator for use over a span of M months from the benchmark month k=1.

   E5: $t_k(S) = (k/M)(\text{HT: } t_k(S)) + (1 - k/M))(\text{E2: } t_k(S))$ for $k = 2, 3, \text{---} M$.

These weights initially (for small k) give most of the weight to E2 and as k increases the weight shifts to HT.

IV. SOME THEORY ON THESE ESTIMATORS

There is a proof in (2) which shows that E0: $t_k(S)$ is model unbiased with respect to (2.1). This means that $E(E0: t_k(S) - Y_k(S)) = 0$ where $Y_k(S) = \sum_{i \in S} Y_k(i)$.

A very similar argument can be used to show that E1 through E3 and R1 and R2 are also model unbiased. As an example, the proof of unbiasedness for E2: $t_k(S)$ follows.

$$E(t_k(S) - Y_k(S)) = E[\, E(t_k(S) - Y_k(S)/k-1)\,]$$

$$= E\,[E(t_k(s_k) - Y_k(s_k)/k-1) +$$

$$E\,((t_k(s_k s_{k-1})/t_{k-1}(s_k s_{k-1}))t_{k-1}(r_k s_{k-1}) - Y_k(r_k s_{k-1})/k-1)$$

$$+\ E\,((t_k(s_k s_{k-2})/t_{k-2}(s_k s_{k-2}))t_{k-2}(r_k r_{k-1}) - Y_k(r_k r_{k-1})/k-1)\,]$$

By definition, the first term in this sum is zero. The second term is equal to:

$$E\,((Y_k(s_k s_{k-1})/t_{k-1}(s_k s_{k-1}))t_{k-1}(r_k s_{k-1}) - Y_k(r_k s_{k-1})/k-1)$$

$$= (\beta_k\, y_{k-1}(s_k s_{k-1})/t_{k-1}(s_k s_{k-1}))t_{k-1}(r_k s_{k-1}) - \beta_k\, y_{k-1}(r_k s_{k-1})$$

$$= (\beta_k\, y_{k-1}(s_k s_{k-1})/y_{k-1}(s_k s_{k-1}))y_{k-1}(r_k s_{k-1}) - \beta_k\, y_{k-1}(r_k s_{k-1})$$

$$= \beta_k\, y_{k-1}(r_k s_{k-1}) - \beta_k\, y_{k-1}(r_k s_{k-1}) = 0$$

$\therefore$ $E(t_k(S) - Y_k(S)) =$ third term $=$

$$\beta_k[E\ (y_{k-1}(s_k s_{k-2})/t_{k-2}(s_k s_{k-2}))t_{k-2}(r_k r_{k-1}) - y_{k-1}(r_k r_{k-1})\,]$$

$$= \beta_k[E\ (Y_{k-1}(s_k s_{k-2})/t_{k-2}(s_k s_{k-2}))t_{k-2}(r_k r_{k-1}) - Y_{k-1}(r_k r_{k-1})\,]$$

Now conditioning on k-2,

$$E\,(Y_{k-1}(s_k s_{k-2})/k-2) = \beta_{k-1}\, y_{k-2}\,(s_k s_{k-2})$$

$\therefore$ $E\,(t_k(S) - Y_k(S)) =$

$$\beta_k\, E(E\,((Y_{k-1}(s_k s_{k-2})/t_{k-2}(s_k s_{k-2}))t_{k-2}(r_k r_{k-1}) - Y_{k-1}(r_k r_{k-1})/k-2))$$

$$= \beta_k\, \beta_{k-1}\, E((y_{k-2}(s_k s_{k-2})/y_{k-2}(s_k s_{k-2}))t_{k-2}(r_k r_{k-1}) - y_{k-2}(r_k r_{k-1}))$$

$$= \beta_k\, \beta_{k-1}[E\ t_{k-2}(r_k r_{k-1}) - Y_{k-2}(r_k r_{k-1})]$$

It has probably been noted by this point that the random variable t and it's realization are used interchangeably. Just which one t represents should be clear from the level of conditioning. The entire preceeding agrument can be repeated with

$r_k r_{k-1}$ in place of S to show that $E(t_{k-2}(r_k r_{k-1}) - Y_{k-2}(r_k r_{k-1})) = \beta_{k-2}\, \beta_{k-3}\, E\,(t_{k-4}\,(r_k r_{k-1} r_{k-2} r_{k-3}) - Y_{k-4}\,(r_k r_{k-1} r_{k-2} r_{k-3}))$. Continuing in this fashion we finally get:

$$E\ (t_k(S) - Y_k(S)) = \prod_{i=0}^{P} \beta_{k-(2i)} \beta_{k-(2i+1)}$$

$$X\ E\,[t_{k-2P}(\Pi_p) - Y_{k-2P}(\Pi_p)\,]$$

where $\Pi_p = \bigcap_{i=0}^{P} r_{k-(2i)}\ r_{k-(2i+1)}$

now for some p, k-2p = 1 or 2. If k-2p = 1 then $E\,[t_1(\Pi_p) - Y_1(\Pi_p)] = 0$ by definition.

If k-2p = 2 then use the model unbiasedness of E1 to see that $E(t_k(S) - Y_k(S)) = 0$.

V. DESCRIPTION OF SIMULATION

These estimators (and combinations of them) were tested on a universe of 300 units each having data for 20 months. A sample of 52 units was selected according to an optimum stratified simple random sampling plan. This sampling plan was optimal with respect to the data for each unit at month one. The data for month one were generated from a lognormal distribution.

The density function for the lognormal is:

$$f(x) = (1/((ax - b)\sqrt{2\pi}))EXP\,(-(1/2a^2) \cdot LOG^2((x - b)/c))$$

where $x > b$

References (3) (4) and (5) indicate that certain economic data are described by the lognormal. For the constants a b and c, I used the following constants which are maximum likelihood estimates derived from an employment survey data set.

$$a = 1.392, \qquad b = .5, \qquad c = 3.158.$$

Thus 300 random numbers were generated from this density for the initial month's universe data, $y_1(i)$. The $y_k(i)$ for k = 2, 3, ---- 20 and i = 1, 2, ---- 300 were generated from the following model.

$$y_k(i) = \beta_k\, y_{k-1}(i) + N(0, y_{k-1}^{2D}(i) \cdot C) + B$$

where $N(0, y_{k-1}^{2D}(i) \cdot C)$ is the normal distribution with mean 0 and variance $y_{k-1}^{2D}(i) \cdot C$.

The triples, (B, C, D), of constants were choosen in a variety of ways and complete simulations were run for each set of triples. This gave some indication of robustness in the eight estimators. For example, a non zero B gives a deviation from the model (1) and the noise level increases with C and D.

The sample of 52 units was selected at month one and used for estimation at each subsequent month.

This stratified simple random sample was choosen as follows.

Nine strata were defined using the cumulative $\sqrt{f}$ rule (6) on the set $y_1(i)$ i = 1, 2, ... 300. Then Neymann allocation was used (6) to divide the sample size of 52 among the strata. Finally a simple random sample was selected in each stratum. This stratified sample was the basis for the Horwitz-Thompson estimator that was mentioned in section III. This sample was also used for the other estimators.

Nonresponse was added to these simulations as follows. The same 52 sample units were kept throughout the 20 months, but a simple random subsample of respondents were selected each month from among the 52 sample units. This occasionally resulted in no respondents in some of the sparser strata so, for the Horwitz-Thompson estimator, strata were collapsed and nonresponse was adjusted by unit counts of respondents and sample members.

In summary, these estimators were tested on a universe of 300 units each having data for 20 months. A sample of 52 units was selected by an optimal stratified simple random sampling plan. The set of respondents, s_k, is a simple random subsample of these 52 units which is selected at each time k. From these samples, s_k, the estimators were computed for each of the 20 months. This procedure was then replicated 20 times and the results were compared to the true population totals at each time k. Thus the mean square error was estimated for each estimator and used as the basis for comparison.

This procedure was repeated for a variety of models. These simulations were done with a response rate of 70% and the set of mixed βs.

VI. TABULAR AND GRAPHICAL RESULTS

Tables 1,2, and 3 show the estimated mean square error in thousands for each estimator studied in a variety of models (B,C,D). These estimated mean square errors are the average over 20 replications and 20 months of the squared deviations between the actual population total and its estimators. B=0 in each of these tables (i.e. the model (2.1) holds). D=0 in table 1, D=.5 in table 2 and D=.7 in table 3. C is the column variable in each of these tables and varies from .1 to .8 in table 1 and from .1 to .4 in tables 2 and 3.

TABLE 1

MODEL (0,C,0)

ESTIMATED TOTAL MEAN SQUARE

ERROR IN THOUSANDS

C	.1	.2	.3	.4	.5	.6	.7	.8
EST								
E 0	0.6	1.3	1.4	3.0	2.3	3.4	3.6	3.9
E 1	0.4	0.9	1.0	1.8	1.7	2.4	2.4	2.6
E 2	0.4	0.9	1.3	1.4	1.8	2.4	2.1	2.8
E 3	0.3	0.9	1.0	1.6	1.9	2.4	2.4	2.8
E 5	9.3	13.8	19.4	26.7	21.6	34.5	38.8	39.3
HT	25.6	31.7	41.5	56.1	48.9	66.7	81.5	85.1
R 1	0.3	0.8	0.9	1.2	1.4	1.8	2.5	2.3
R 2	0.4	1.0	1.3	1.5	2.1	2.6	2.6	2.3

The first thing that should be noted is that when D=0, $V_k(i)$ is independent of $x_{k-1}(i)$, then R1 generally outperforms the other estimators. This is exactly what one would expect since if D=0 then R1, based on d=0, is the estimator suggested by least squares regression theory outlined in Section II. Table 1 also shows that E1 through E3 do nearly as well as R1 when D=0. Recall that the regression theory of Section II suggests these estimators when D=.5.

TABLE 2

MODEL (0,C,.5)

ESTIMATED TOTAL MEAN SQUARE

ERROR IN THOUSANDS

C	.1	.2	.3	.4
EST				
E0	13.5	34.5	45.1	76.1
E1	7.2	17.7	26.3	38.3
E2	7.9	18.0	28.0	33.0
E3	7.1	17.1	25.8	36.5
E5	39.8	67.5	143.7	193.2
HT	83.0	121.2	251.1	339.1
R1	12.3	25.4	32.0	42.2
R2	8.9	19.8	29.0	40.4

When D=.5 (table 2), then E1 through E3, based on d=1, dominate as the theory would suggest. R1 does not do particularly well in this case however. Tables 1 through 3 suggest that the estimates based on the assumption that d=1 are quite robust for large deviations from d=1 (i.e. from D=.5) and that R1, based on d=0, does not seem to share this property.

Next note that HT, E5 and E0 are not competitive when any of the models that are tabulated in tables 1 through 3 hold. R2 does reasonably well but seems to be slightly inferior to E1, E2 and E3.

TABLE 3

MODEL (0,C,.7)

ESTIMATED TOTAL MEAN SQUARE

ERROR IN THOUSANDS

C	.1	.2	.3	.4
EST				
E0	63.2	174.5	226.5	387.9
E1	32.4	86.3	126.4	193.3
E2	34.1	78.8	121.3	173.5
E3	35.4	80.3	122.9	189.4
E5	108.0	168.6	431.9	652.6
HT	198.7	282.5	707.3	1028.0
R1	78.1	157.5	193.9	277.9
R2	43.0	86.7	133.3	216.8

Figure 1 is a graph by month of the estimated mean square error in hundreds from 20 replications for three estimators E0, E1 and E2. It begins at month 4 and goes to month 20. In this example it is clear that E1 is very nearly a two fold improvement over the link relative estimator, E0, with respect to mean square error. Although E2 has a lower average mean square error than E1, it develops a disturbing oscillation in latter months. This behavior is common with E2 in other models and the reason for this is not yet known.

FIGURE 1

MODEL (0, .4, .5)

ESTIMATED MEAN SQUARE ERROR

BY MONTH IN HUNDREDS

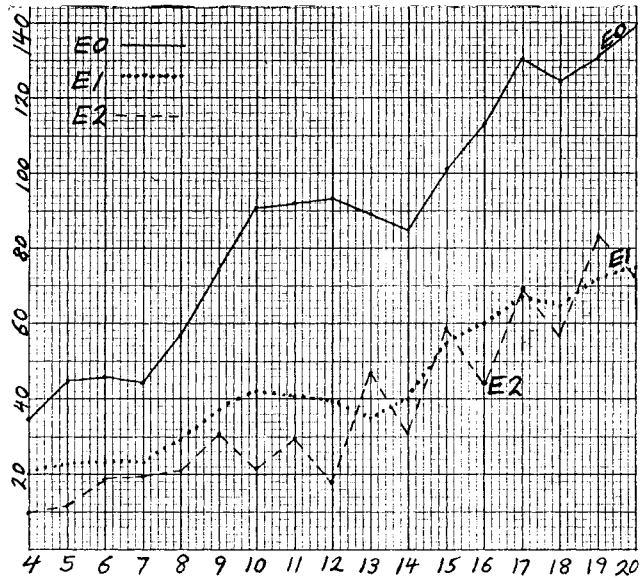


Figure 1 also shows the diminishing return in the use of historical data. E2 is better than E1 but the improvement is not very substantial. Compare this improvement to the difference between E1 and E0.

Tables 4 and 5 are summary tables that show which estimator performed best in a wider variety of models. Table 4 runs through combinations of D and C with B=0. The model (2.1) holds in all these cases (B=0). The conditional variance structure is a function of C&D. When D=0 the theory suggests R1 should do best and table 4 shows clearly that this is the case. It also shows that for any value of D near .5 (d near 1) then the E type estimators are quite robust. It should also be noted that when D=.5, E3 dominates but when D=.3 or .7, E1 or E2 usually dominate. This behavior is also predicted by the theory in Section II where it is shown that for $d \neq 1$ the estimation of changes between non adjacent months is a difficult problem which is not handled optimally by the E type estimators. Thus the estimator which uses the most historic data (E3) does not hold up well when d deviates from 1 because it is not using an optimum update factor for linking forward from month k-2 to month k.

The estimators which use less historic data (E1 & E2) suffer less from this problem when d differs from 1.

Table 5 shows what happens if an intercept term is added to the model (that is $B \neq 0$). In this case B=.1. When this happens the composite estimator E5 does quite well for small values of D and C. As D and C become larger then the situation that B=.1 contributes proportionally less than C&D to the overall between month change and as a result the regression estimators based on d=1 dominate.

TABLE 4

MODEL (0,C,D)

BEST ESTIMATOR BY MODEL

D/C	.1	.2	.3	.4
0	R1/E3 *	R1	R1	R1
.3	E3	E1	E3	E2
.5	E3	E3	E3	E2
.7	E1	E2	E2	E2

*Tie between R1 & E3.

If d=1 then E type estimators (E1, E2, E3) are good with E3 slightly better than E1. If $d \neq 1$ then the use of more than 2 months of data (k,k-1) is hazardous because of the problem of estimating products of β's. Thus if d is unknown, which it is for practical purposes, then the E type estimator using 2 months of data (k,k-1) is suggested. These facts lead to the use of E1 because it uses only 2 months of data but remains robust when d deviates from 1.

TABLE 5

MODEL (.1,C,D)

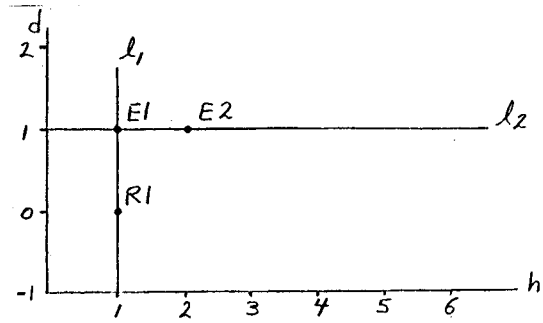
BEST ESTIMATOR BY MODEL

D/C	.1	.2	.3	.4
0	E5	E5	E5	E5
.3	E5	E5	R2	E2
.5	R2	E3	R2	E2
.7	R2	E3	E2	E2

VI. CONCLUSIONS

Before the conclusions are stated the set of estimators that were considered will be reviewed. Except for the Horwitz-Thompson estimator and the composite estimator that contains the Horwitz-Thompson estimator, all the others can be classified according to points of the two dimensional array, (d,h), that was mentioned in the Introduction.

If the horizontal axis indicates the number of months of data used by a particular estimator and the vertical axis indicates the value of d then the estimators that were considered correspond to points in the following graph:



As an example consider the point labeled E2 in the table. E2 is the regression estimator that uses data from times k, k-1 and k-2, 3 times periods, and is based on d=1.

When d is not one it is difficult to use a regression estimator that uses more than two months of data (k, k-1) because of the difficulty in estimating products of β's. This problem was pointed out in Section II. If an estimator such as E2 or E3 is used when d is not one, the tabular results indicate that more is lost when products of β's are incorrectly estimated than is gained by the use of data from months k-2 and k-3. For practical purposes d is never known exactly. Thus for reasons of both accuracy and simplicity it is recommended that an estimator which uses data from only time k and k-1 be used. This narrows down the set of estimators to those corresponding to points that lie on l_1, in the graph.

It may be possible to estimate d and although this estimator, $\hat{d}$, may not be very good, a regression estimator based on $\hat{d}$ and using data from times k and k-1 is recommended. If d is not estimated then a good rule of thumb, as shown by the computer simulations, is to choose E1. E1 is the regression estimator based on d=1 and remains quite robust for large deviations of the true value of d from one.

The monthly survey of employment, hours, and earnings at the Bureau of Labor Statistics is undergoing revision. The work that is summarized in this paper is part of this revision. E1 is a simple modification of E0, the estimator that is currently being used in the this survey. This modification can be described by saying that E1 lets the sample of respondents represent itself. To show explicitly what this means recall that E0 can be written

$$E0:\ t_k(S) = (y_k(s_k s_{k-1})/y_{k-1}(s_k s_{k-1}))t_{k-1}(s_k) + (y_k(s_k s_{k-1})/y_{k-1}(s_k s_{k-1}))t_{k-1}(r_k)$$

E1 replaces the first term of this sum with the observed sample values at time k.

$$E1:\ t_k(S) = y_k(s_k) + (y_k(s_k s_{k-1})/y_{k-1}(s_k s_{k-1}))t_{k-1}(r_k)$$

This simple modification of the current system should be a good choice among the possible estimators studied in this paper.

E1, like the other estimators that correspond to points of l_1 or l_2 is unbiased with respect to model (2.1).

The problem of estimating variances for the estimators considered here has yet to be looked at.

My thanks are due to Wesley L. Schaible, Alfreda Reeves, Sandra West and Bennett Brady for their help in preparing this thing.

REFERENCES

1) Madow, Lillian H., Madow, William G. (1978) "On Link Relative Estimators", ASA Proceedings of the Section on Survey research Methods, 534-9.

2) Royall, Richard M. (1981), "Study of Role of Probability Models in 790 Survey Design and Estimation", Bureau of Labor Statistics contract report 80-98.

3) Johnson, N.L., Kotz, S. (1970), "Continuous Univariate Distributions -I", John Wiley and Sons, New York, Chapter 14.

4) Gibrat, R. (1930), "Une loi des reportitions economiques: l'effet proportionnel", Bulletion de Statistique General, France 19, 469ff.

5) Gibrat, R. (1931), "Les Ineqalities Economiques, Paris: Libraire du Recueil Sirey.

6) Cochran, William (1963), "Sampling Techniques", Second Edition, John Wiley and Sons, New York, pp. 95 and pp. 130.