IMPUTATION IN A PERIODIC SURVEY

Phillip S. Kott, Energy Info. Adm.

The problem of how to impute for nonresponse in a survey conducted repeatedly over time has seen little theoretical development. This paper provides a model-based analysis of the updated historical imputation strategy - an approach to imputation dubbed the "ratio-of-identicals method" by the Census Bureau (for example, see Huang 1984).

While in practice surveys are often considerably more complicated than anything discussed here, the analysis does bear edible fruit. Key conceptual issues are isolated and a practical test for evaluating alternative imputation strategies is introduced.

Section 1 discusses the standard technique of imputation with the respondent mean (of a cell) both in terms of a parametric and a quasi-random response model. Section 2 develops the updated historical imputation methodology and investigates its properties under these two types of models. Section 3 proposes a times series model under which there are potential gains from exponentially smoothing historical values. A test is introduced in section 4 for comparing alternative mechanisms for calculating the historical values.
An empirical example using monthly gasoline volumes reported to the Energy Information Administration follows.

1. THE STANDARD FRAMEWORK

1.1 The Problem

Consider a survey conducted among a population of N units to estimate the total quantity of some parameter of interest. Let X_i be the quantity of the parameter contained by unit i, and $X=\sum^N X_i$.

We will assume that, in the absense of nonresponse to the survey, X is estimated based on a simple random sample of $n \leq N$ distinct units. We allow the possibility that $n=N$; in other words, the survey may be a complete census.

We are restricting the theoretical analysis in this and subsequent sections to a unstratified population, but it is possible to think of the population under study as a single stratum or cell of a larger population. In the example offered in Section 4, this is indeed the case.

One estimator of X is the simple expansion estimator: $\hat{X}_E=(N/n)\sum^n X_i$. Throughout the paper, units are relabelled so that $\sum^k Y_i$ sums only the Y-values of those units in the relevant class of k units. In this case, $\sum^n X_i$ is the sum of the parameter of interest contained by the n units in the sample exclusively.

Now suppose that among the n units in the sample only n_R units respond to the survey. If that is the case, one must impute values for the remaining n_{NR} units.

1.2 A Simple Imputation Strategy

The simplest and most popular imputation strategy is to proxy the missing value of each nonrespondent by the average value of the respondents. In our framework, this means to use $X^*=\sum^{n_R} X_i/n_R$ in place of the nonrespondent values in $\hat{X}_E$:

$$\begin{aligned}\hat{X}_E &= (N/n)\sum^n X_i \\ &= (N/n)\left[\sum^{n_R} X_i + \sum^{n_{NR}} X^*\right] \\ &= (N/n_R)\sum^{n_R} X_i \,.\end{aligned} \tag{1}$$

The estimator in the last line of (1) has the same form as the expansion estimator with n_R replacing n. As a result, we will also denote it as $\hat{X}_E$. This should not cause any confusion.

1.3 A Parametric Model

Implicit in the development of $\hat{X}_E$ in the last subsection is the assumption that nonresponding units are similar to respondents. One way to formalize this similarity is by a parametric model in which every unit has the form:

$$X_i = \bar{\mu}(1+\varepsilon_i) \tag{2}$$

where ε_i is a random variable with mean zero. Equation (2) stipulates that the differences among the X_i can be treated as random noise. (N.B. In this paper, we have abstracted away from stratified sampling designs, where one has the luxury of applying a different version of (2) to each stratum.)

It is easy to see that under the model in (2), $\hat{X}_E$ is an unbiased estimator:

$$E_M(\hat{X}_E - X) = N\bar{\mu} - N\bar{\mu} = 0 .$$

(The subscript "M" on the expectation operator is used to specify that the expectation is with respect to the model.)

If the ε_i are independent and identically distributed, then $\hat{X}_E$ is the best linear unbiased estimator of X given only the n_R responses. This is well known. Suppose, however, a survey of X-values is taken repeatedly over time, and some units that have failed to respond to the current survey did respond to a previous survey. If that is the case, $\hat{X}_E$ may be improved upon by using the information contained in the previous survey. More on this in a later section.

1.4 A Response Model

Many survey statisticians are uncomfortable with the parametric model expressed in equation (2). They would prefer to estimate X free of any assumptions about the parameter of interest. Assumptions can, after all, be wrong. An assumption-free approach is possible, however, only in the absense of nonresponse. When faced with the spectre of nonresponse, these statisticians are forced to use a model. The model they use, however, is of response behavior rather than parametric behavior.

In a quasi-randomized response model, nonresponse is treated as the realization of a random variable. Each unit is assumed to have a positive probability of responding to the survey. Response probabilities become little more than another layer of the random selection process in the design-based theory of sample design and inference.

The simple quasi-random response model we will use here assumes that each unit is equally likely to repond to the survey. It is then possible to show that the expectation of $\hat{X}_E$ with repect to the survey design (simple random sampling without replacement) and the response model is X. As a result, we say that $\hat{X}_E$ is design unbiased. (Since we have defined design unbiasedness with respect to a response model, $\hat{X}_E$ is said to be design unbiased even when n=N, and there is no sampling design except in a trivial sense.)

The design unbiasedness of $\hat{X}_E$ is, strictly speaking, conditional on the number of respondents being positive ($n_R>0$). For a more thorough introduction to the design-based theory of imputation complete with a proof of the conditional design unbiasedness of $\hat{X}_E$, the reader is referred to Oh and Scheuren (1983).

In practice, it is rarely the case that all units are equally likely to respond. Statisticians are aware of this and attempt to partition the population into response classes containing units with equal probabilities of nonresponse. For our purposes, it is useful to abstract away from the need to break up the population into response classes just as we abstracted away from complicated survey designs.

One last point and we will be ready to tackle updated historical imputation. Recall that $\hat{X}_E$ is design unbiased under the response model no matter how the X_i are specifed. In a similar vein, observe that $\hat{X}_E$ is model unbiased under (2) even if the units have different likelihoods of response (or different probabilities of selection for that matter).

2. THE UPDATED HISTORICAL VALUE

2.1 The Methodology

Suppose previous X-values for a nonrespondent i are known. One reasonable method for imputing X_i is to "update the historical value" of unit i:

$$X_i^* = \frac{\sum^{n_R} X_j}{\sum^{n_R} \tilde{X}_j} \tilde{X}_i , \tag{3}$$

where X_i^* is the imputed value of i, and
$\tilde{X}_k$ is the historical value of k, which is derived from k's (and perhaps some other units') previously reported values (the mechanism for determining $\tilde{X}_k$ will be left vague for the moment).

Equation (3) says that a proxy for X_i is found by multiplying the historical value of i, $\tilde{X}_i$, by an estimate for the ratio of i's current value to its historical value. We will call this ratio, $X_i/\tilde{X}_i$, the growth rate of i. The estimate of this rate in (3), the so called "ratio-of-identicals," is simply the (weighted) average of the growth rates of the respondents. Using this proxy, the estimator of X is

$$\begin{aligned} \hat{X}_U &= (N/n) \sum X_i \\ &= (N/n)\left[\sum^{n_R} X_i + n_{NR} X_i^*\right] \\ &= (N/n) \sum \tilde{X}_i \left(\sum^{n_R} X_j / \sum^{n_R} \tilde{X}_j\right) . \end{aligned} \tag{4}$$

2.2 A Parametric Model

The imputation strategy expressed by equations (3) and (4) can be justified using the following two equation "random effects" model:

$$X_i = \beta' \tilde{X}_i (1+\varepsilon_{2i}') \tag{5}$$

$$\tilde{X}_i = \mu(1+\varepsilon_{1i}), \tag{6}$$

where ε_{1i} and ε_{2i}' have means of zero and are respectively independent across units (e.g., $E(\varepsilon_{ki}'\varepsilon_{kj}')=0$ for k=1 or 2, $i \neq j$). It is not necessary for ε_{1i} and ε_{2i}' to be uncorrelated. Nor is it necessary for the ε_{1i} to be identically distributed given a vector of $\tilde{X}_i$ values. We do assume, however, that they are identically distributed unconditionally, that is, before the values of the ε_{1i} (and thus the $\tilde{X}_i$) become known. Moreover, we assume that the ε_{1i} are identically distributed and that the covariance of ε_{2i}' and ε_{1i} is constant over the units.

What equations (5) and (6) say in words is that the differences among the X_i have two random sources. Source one is the differences among the historical values. The deviation of $\tilde{X}_i$ from the common mean is the random variable $\mu\varepsilon_{1i}$. Source two is the differences among the unit growth rates. The deviation of $X_i/\tilde{X}_i$ from the common mean is $\beta'\varepsilon_{2i}'$.

The imputation strategy in (3) captures the first source of deviation contained in the unknown X_i, but not the second. It does this by employing the historical value, $\tilde{X}_i$, in determining $\hat{X}_U$. On the other hand, the strategy of imputing using the respondent mean fails to capture either source of deviation. Intuitively, unless there is strong dependency between the two random components, the updated historical imputation methodology should prove to be superior.

2.3 A Theorem

It is helpful to recast the structural equations (5) and (6) into a reduced form:

$$X_i = \beta'\mu(1+\varepsilon_{1i} + \varepsilon_{2i}' + \varepsilon_{1i}\varepsilon_{2i}') . \tag{7}$$

It is a simple exercise (see the appendix, which is available from the author upon request) to redefine β and ε_{2i} so that

$$X_i = \beta\mu(1+\varepsilon_{1i} + \varepsilon_{2i}), \tag{8}$$

where the mean of ε_{2i} is zero. While β and ε_{2i} do not have the simple intuitive interpretations of their original counterparts, the subsequent mathematics is streamlined with their use.

In the Appendix, the following lemma is proven.

Lemma 1. If equation (8) holds with

$E(\varepsilon_{ki})=0$, for k=1,2, all i; (9.1)

$E(g_i g_j)=E(g_i)E(g_j)$
for $g_k=g(\varepsilon_{1k},\varepsilon_{2k})$, $i \neq j$ (9.2)

$E(\varepsilon_{ki}^2)=\sigma_k^2$, for k=1,2; (9.3)

$E(\varepsilon_{1i}\varepsilon_{2i})=\rho\sigma_1\sigma_2$; (9.4)

$1/n_R^2 \doteq 0$, (9.5)

then

$$E_M(\hat{X}_E - X)=0, \tag{10}$$

$$E_M[(\hat{X}_E - X)^2] = \beta^2\mu^2N^2(\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2)\left(\frac{1}{n_R}-\frac{1}{N}\right) \tag{11}$$

$$E_M(\hat{X}_U - X) \doteq -\beta\mu N\frac{n_{NR}}{n_R}\rho\sigma_1\sigma_2 \tag{12}$$

and

$$E_M[(\hat{X}_U - X)^2] \doteq \beta^2\mu^2N^2\left[\sigma_2^2\left(\frac{1}{n_R}-\frac{1}{N}\right)+(2\rho\sigma_1\sigma_2+\sigma_1^2)\left(\frac{1}{n}-\frac{1}{N}\right)\right]. \tag{13}$$

(N.B. As was the case with ε_{2i}', the ε_{2i} need only have identical variances unconditionally.)

Equation (13) tells us that $\hat{X}_U$ is not model unbiased unless ε_{1i} and ε_{2i} are uncorrelated. Nevertheless, from (11) and (13), one can see that the model mean squared error of $\hat{X}_U$ is no more than the model mean squared error of $\hat{X}_E$ as long as ε_{1i} and ε_{2i} are not so inversely correlated that $\rho < -\sigma_1/(2\sigma_2)$. Let us state this more precisely as a theorem:

Theorem 1. If equations (6), (8), and (9) all hold, and $n_R<n$, then
$MSE(\hat{X}_U) \lesseqgtr MSE(\hat{X}_E)$ when $\rho \gtreqless -\sigma_1/(2\sigma_2)$.

The theorem tells us that under the model and assumptions (9.1)-(9.5), the updated historical imputation methodology will be more efficient than the respondent mean methodology except in some applications (but not all) where the unit growth rates are inversely correlated with the historical values. In Section 3, a reason for such inverse correlation will be offerred as well as a practical remedy.

A useful alternative to (13) is

$$E_M[(\hat{X}_U - X)^2] \doteq \beta^2\mu^2N^2(\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2)\left(\frac{1}{n}-\frac{1}{N}\right)+ \beta^2\mu^2N^2\sigma_2^2\left(\frac{1}{n_R}-\frac{1}{n}\right). \tag{14}$$

Also of future use to us is the fact that in the degenerate case where all the unit historical values are equal (say to unity), $\sigma_1^2=0$, and $\hat{X}_U$ collapses into $\hat{X}_E$.

2.4 Design Consistency

The model expressed by equations (5) and (6) is very simplistic. Building an imputation strategy solely on this parametric model may produce an unwanted systematic bias in certain applications.

A form of protection against parametric model misspecification is design consistency (Isaki and Fuller, 1982). Given the quasi-random response model discussed in Subsection 1.4, design consistency requires that $(\hat{X}_U-X)/X$ converges to zero in design probability as the sample size, n, grows arbitrarily large. Note that "design probability" is defined with repect to both the sampling design and the response model.

For any asymptotic property to be demonstrated, certain boundary assumptions are needed. To show that $\hat{X}_U$ is design consistent, these assumptions are sufficient as n tends toward infinity:

$\text{plim}_D\; n_R/n=A>0$,

$\lim \sum^N X_i/N=B<\infty$,

$\lim \sum^N \tilde{X}_i/N=C<\infty$,

$\lim \sum^N (X_i-B)^2/N=D<\infty$, and

$\lim \sum^N (\tilde{X}_i-C)^2/N=E<\infty$.

These assumptions assure that $F=(\sum^{n_R} X_i/n_R - \sum X_i/N)$ converges in design probability to zero (its design exectation is zero, while it design variance, $(1/n_R - 1/N)D$, is of order n^{-1}). Similarly, $(\sum \tilde{X}_i/n)/(\sum^{n_R}\tilde{X}_i/n_R)$ converges in design probability to one. Thus, noting from the last line of (4) that $\hat{X}_U=(N/n)\sum \tilde{X}_i(\sum^{n_R} X_j/\sum^{n_R}\tilde{X}_j)$, it is clear $(\hat{X}_U-X)/X$ converges in design probability to zero.

The estimator $\hat{X}_U$ may not be the most efficient under the parametric model in equations (5) and (6). Huang (1985) investigates an alternative estimator, $\hat{X}_U^*$, that imputes for a missing X_i with the formula $X_i^*=\tilde{X}_i\sum^{n_R}(X_j/\tilde{X}_j)/n_R$ rather than with equation (3). This estimator is model efficient if the ε_{2i}' in (5) are identically distributed given the $\tilde{X}_i$ (ε_{1i}). Using real data and simulating nonresponse by a simple random process, Huang found that $\hat{X}_U^*$ appeared to be more biased than $\hat{X}_U$. This is a direct consequence of $\hat{X}_U$, but not $\hat{X}_U^*$, being design consistent under the simple random response model she simulated (the same one we have been assuming). Design consistency itself depends on the validity of the response model. Ironically, the parametric model provides some protection against the possibility of response model failure. If both the parametric and the response model are misspecified, however, the imputation strategy will be flawed. (It is thus prudent to try to stratify the population in such a way that both models hold or nearly hold in every cell.)

3. EXPONENTIAL SMOOTHING

3.1 The Need

Up until now, we have discussed the concepts of an historical value and a repeated survey only in vague terms. Let us tighten them up a bit. Suppose surveys are conducted at equidistant time periods denoted 0, 1, ..., T. Let X_{it} be the X-value of unit i at time t, and $\tilde{X}_{it} = X_{it-1}$ (assuming X_{it-1} is known). In other words, the historical value for a unit at a particular time is simply its value in the previous period.

This is a common formulation of the historical value in practice. It has two drawbacks. The first is that X_{it-1} might not be known. This situation is easily handled by "moving forward" the last reported X_i-value, say X_{ir}, by the "linked-average" growth rate of respondents since then; i.e.,

$$\tilde{X}_{it} = \left(\frac{\sum_{j\in R^{t-1}} X_{jt-1}}{\sum_{j\in R^{t-1}} \tilde{X}_{jt-1}}\right)\cdots\left(\frac{\sum_{j\in R^{r+1}} X_{jr+1}}{\sum_{j\in R^{r+1}} \tilde{X}_{jr+1}}\right) X_{ir},$$

where R^u is the set of units responding in period u. Note that $\tilde{X}_{is}$, s<t, when X_{is-1} is unknown is defined recursively.

The other drawback of using last period's response for this period's historical value is that a given unit's response in a particular period may be a temporary abberation. As a result, a higher than expected X_{it-1} value will often be followed by a lower than expected X_{it} value. Mathematically, if $X_{it-1} > \beta_{t-1}X_{it-2}$, then $X_{it} < \beta_t X_{it-1}$ will often obtain, where β_s is the model growth rate from s-1 to s.

The upshot of this is a tendency for the growth rate of a unit, with $\tilde{X}_{it}$ defined as X_{it-1} (X_{it-1} known), to be inversely correlated with its historical value. This is precisely the situation warned against in Subsection 2.3!

3.2 The Model

The remedy for the possibilty of a one period abberation in a unit's value is to <u>exponentially</u> <u>smooth</u> the historical value, This procedure has been developed in the time series literature (for example, see Fuller, 1976). In the context of imputation it takes on a somewhat special form as we shall see.

Consider this model:

$$X_{it} = \beta_t X_{it-1} + \eta_{it} - \lambda\beta_t\eta_{it-1},$$

where t=1,...T;
$0 \le \lambda \le 1$; and

$E(\eta_{is}\eta_{ju}) = 0$ for $i \ne j$ or $s \ne u$.

If $\lambda > 0$, the model builds in a probability that one period increases (or decreases) in X_{is} will be reversed in the subsequent period. It does this by hypothesizing that the disturbances obey a first order autoregressive process.

To make matters simpler, let $X_{i0} = \mu_{i0}$. This assumes that the initial value, X_{i0}, is not itself a one period abberation. Observe that $E_M(X_{it}) = \mu_{it} = \prod_{s=1}^{t}\beta_s\mu_{i0}$.

The final assumption of this time series model is that the variance of X_{it} is proportional to the square of its mean. This makes

$$X_{it} = \beta_t X_{it-1} + \mu_{it}(\varepsilon_{it} - \lambda\varepsilon_{it-1}), \quad (15)$$

where $Var(\varepsilon_{is}) = \sigma_i^2$. (N.B. In this section, and this section only, we have the luxury of allowing the $Var(\varepsilon_{is})$ to vary across the units for a particular s.)

We can relabel X_{it}/μ_{it} as y_{it} to get the simple model:

$$y_{it} = y_{it-1} + \varepsilon_{it} - \lambda\varepsilon_{it-1}. \quad (16)$$

Equation (16) can be rewritten in serially independent (over time) and homoskedastic (ditto) form as

$$y_{it} = (1-\lambda)y_{it-1} + (1-\lambda)\lambda y_{it-2} + (1-\lambda)\lambda^2 y_{it-3} + \ldots + \lambda^{t-1}y_{i0} + \varepsilon_{it}.$$

In terms of the X_{it}, this is

$$X_{it} = \beta_t\left[(1-\lambda)X_{it-1} + \ldots + \prod^{t-1}\beta^s X_{i0}\right] + \mu_{it}\varepsilon_{it}. \quad (17)$$

The bracketed expression in (17) is the <u>exponentially</u> <u>smoothed</u> historical value of unit i, where λ is the <u>smoothing</u> <u>parameter</u>. When λ is zero, no smoothing takes place. As λ increases, this $\tilde{X}_{it}$ becomes less a function of X_{it-1} and more a function of unit i's previous X-values.

By using a smoothed historical value in the imputation formula in equation (4) we remove <u>at least part</u> of the tendency for the growth rate and the historical value to be inversely related (perhaps only part, because the ε_{it} may yet be negatively correlated with the X_{i0}). This reduces some, if not all, of the model bias of $\hat{X}_U$ (see equation (12)). In addition, it stands to reason that since the model variance of X_{it} in (17) conditional on $\tilde{X}_{it}$ is less than that of X_{it} in (15) conditional on X_{it-1}, the model mean squared error of $\hat{X}_U$ is less when the smoothed historical value is used in place of last period's value. The exact link between the conditional variance of X_{it} and the model mean squared error of $\hat{X}_U$ will be established in Section 4.

3.3 Estimation

The smoothed historical value in recursive form is

$$\tilde{X}_{it} = (1-\lambda)X_{it-1} + \lambda\beta_{t-1}\tilde{X}_{it-1}$$

In the context of a stationary stochastic process for which exponentially smoothing was developed, all the β_s, $s \le t-1$, are unity, If that were the case here, one could aggregate the X_{it} over the units in some manner, and then estimate λ from a time series using an ARIMA(0,1,1) package. (ARIMA stands for Auto-Regressive Integrated Moving Average. An ARIMA(0,1,1) model is simply an integrated moving average of order one. Equation (16) is an example.) In most survey applications, however, the X-values are seasonal or trending. As a result, the β_s can not be reasonably treated as if they were all one.

Fortunately, one does not have to assume anything about the β_s. They can be circumnavigated by seperating the units into two groups, G_1 and G_2 ; letting $X_t^1 = \sum_{i\in G_1} X_{it}$, and $X_t^2 = \sum_{i\in G_2} X_{it}$; and running $Y_t = X_t^1/X_t^2$ through an ARIMA(0,1,1) package. It is a tedious but straight forward exercise to show that Y_t approximately obeys an ARIMA(0,1,1) model with parameter when each X_{it} obeys (15).

Most often in practice, one will not have the luxury of a long enough time series to estimate λ from the data with an ARIMA package. Instead an "estimated" λ value, call it θ, must be determined from external sources or indirectly using the test developed in the following section. Armed with any estimate of λ, one still needs to estimate the β_s, $0 < s \le t-1$, before an appropriate $\tilde{X}_{it}$ can be determined.

The β_s can be estimated (recursively) by

$$\hat{\beta}_s = \sum_{j\in R^s} X_{js} / \sum_{j\in R^s} \tilde{X}_{js}$$

The exponentially smoothed historical value, $\tilde{X}_{js}$, for s>1, is determined by

$$\tilde{X}_{js} = (1-\theta)X_{js-1}^{(\sim)} + \theta\hat{\beta}_{s-1}\tilde{X}_{js-1} \quad (18)$$

where $X_{js-1}^{(\sim)}$ is X_{js-1} if X_{js-1} is known and $\hat{\beta}_{s-1}\tilde{X}_{js-1}$ otherwise. For s=1, $\tilde{X}_{js} = X_{j0}$. This requires that X_{j0} must be known for all j.

4. COMPARING ALTERNATIVE HISTORICAL VALUES

4.1 The Test

We now return to the model (and notation for the ε) based on equations (6), (8), and (9), where the mechanism for the determination of the $\tilde{X}_i$ is unspecified. Let $\tilde{\underset{\sim}{X}}$ be the vector $(\tilde{X}_1,\ldots,\tilde{X}_n)$. In this section, $\tilde{X}_i$ need not be an exponentially smoothed version of the X-value in the previous period. It may be any function of previous X-values.

Theorem 2. Suppose that the model in (6), (8) and (9) holds given either of a pair of $\tilde{\underset{\sim}{X}}$ vectors, $\tilde{\underset{\sim}{X}}^1$ and $\tilde{\underset{\sim}{X}}^2$ (these vectors are the results of alternative methods of calculating the $\tilde{X}_i$). Let $G_i(\tilde{\underset{\sim}{X}}^k) = X_i - \beta^k\tilde{X}_i^k$. If the relative variance of of $G_i(\tilde{\underset{\sim}{X}}^1)$ is less than that of $G_i(\tilde{\underset{\sim}{X}}^2)$, then the model mean squared error of $\hat{X}_U$ based on $\tilde{\underset{\sim}{X}}^1$ will be will be less than that of $\hat{X}_U$ based on $\tilde{\underset{\sim}{X}}^2$.

Proving this theorem is a simple matter. Observe that the relative variance of $G_i(\underset{\sim}{X})$ is simply σ_2^2 (see (5) and (8)). From (8), we see that the relative variance of X_i is $\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2$. Since this relative variance is invariant to the choice of $\underset{\sim}{X}$, the model mean squared error of $\hat{X}_v$ expressed (14) can be seen to be a linear function of σ_2^2 - the relative variance of $G_i(\underset{\sim}{X})$. QED.

Given a vector $\underset{\sim}{\tilde{X}}$, σ_2^2 can be estimated by

$$M(\underset{\sim}{\tilde{X}})=\sum^{n_R}\left(x_i-\left[\sum^{n_R}x_j/\sum\tilde{x}_j\right]\tilde{x}_i\right)\Big/(n_R-1)\left(\frac{\sum^{n_R}x_i}{n_R}\right)^2. \quad (19)$$

Under the assumptions in (9), $M(\underset{\sim}{\tilde{X}})$ is (approximately) unbiased.

Equation (19) provides a powerful indirect test for choosing between alternative calculations of the $\underset{\sim}{\tilde{X}}$. While not restricted to historical values of the form in equation (18), this test can nonetheless be used to evaluate different values of smoothing parameter as will be seen shortly.

An interesting corollary to Theorem 2 follows.

Corollary 2.1 Suppose the model in (6), (8) and (9) holds given a vector $\underset{\sim}{\tilde{X}}$. If ρ in (9.4) is non-zero, then there exists a vector $\underset{\sim}{\tilde{X}}^*$ such that the model holds, and $\hat{X}_v$ based on $\underset{\sim}{\tilde{X}}^*$ has less model mean squared error than $\hat{X}_v$ based on $\underset{\sim}{\tilde{X}}$.

The proof of this corollary involves calculating the $\tilde{X}_i^*$ so that $\mu\varepsilon_{1i}^* = \tilde{X}_i^*-\mu = \mu\varepsilon_{1i}(1+\rho\sigma_2/\sigma_1)$. This can always be done in principle. In practice, values for μ, ρ, σ_1, and σ_2 are needed before the $\tilde{X}_i^*$ can be computed. (Note that $\varepsilon_{2i}^*=\varepsilon_{2i}-\varepsilon_{1i}\rho\sigma_2/\sigma_1$, so that $\sigma_2^{*2}=\sigma_2^2(1-\rho^2)$.)

4.2 An Example

The Energy Information Administration (EIA) collects monthly State-level sales volumes and revenues for a variety of petroleum products and uses in its EIA-782 survey (see any issue of The Petroleum Marketing Monthly). In this subsection, our attention is focused on the imputation of March 1985 volumes for survey nonrespondents in the nine gasoline product/use categories. The products are leaded, unleaded, and premium gasoline sold through company owned outlets, to other end users, or for resale. This three by three matrix constitutes the nine product/uses, which will be called simply products from now on.

Reporting units in each of the 50 States and the District of Columbia have been divided by EIA into 10 cells for imputation purposes. Many of the product/State/cells (called from now on cells) are empty. Some have only a few members and must be collapsed into other cells when all the members fail to respond in a given month.

While collapsing poses an interesting question in its own right, it is beyond the scope of this endevor. The empirical analysis discussed here was restricted to units with positive responses in each of the four months between December 1984 and March 1985 and to cells containing at least two such units.

Each of the remaining cells was treated as a population. Seven methods of calculating March 1985 historical values were investigated. The first method set all the historical values equal to unity (this results in imputing with the respondent mean in each cell). The second method used reported February volumes as the historical values. The remaining five methods exponetially smoothed the February volumes with smoothing parameters of .1, .2, .3, .4, and .5 respectively. Historical values were truncated at the December 1984 term. If t is March 1984, then

$$\tilde{X}_{it}(\lambda)=(1-\lambda)X_{it-1}+\hat{\beta}_{t-1}(1-\lambda)\lambda X_{it-2}+\hat{\beta}_{t-1}\hat{\beta}_{t-2}\lambda^2 X_{it-3},$$

for $\lambda=.1, \ldots, .5$.

The test statistic in equation (19) has been computed for each of the seven methods of detemining historical values. For a cell k, we call these test statistics ME_k, ML_k, $MS1_k$, $MS2_k$, $MS3_k$, $MS4_k$, and $MS5_k$ respectively.

Let $\hat{X}_{Ek}$ be the estimated total volume for cell k in March based on using the first imputation strategy, and let $\hat{X}_{Lk}$, $\hat{X}_{S1k}$, etc. be defined conformally. If the model mean squared error of $\hat{X}_{Lk}$ is less than that of $\hat{X}_{Ek}$, we expect ME_k-ML_k to be positive. As a result, when ME_k-ML_k is positive, we say that $\hat{X}_{Lk}$ is more likely the better estimator.

In Table 1 the number of cells in which ME_k-ML_k is positive and nonpositive is displayed for each of the nine products. The difference between the two is signficant when it is greater the twice the square root of their sum (the null hypothesis is the binomial distribution with p=.5; significance is at the .05 level).

For all the products, ME_k-ML_k is positive significantly more often than not. From this we can conclude that for the problem at hand, updated historical imputation using last month's response as the historical value is likely to be better than imputation with the respondent mean in a significant majority of cells.

Also displayed in Table 1 is the number of times $MS1_k-ML_k$ is positive and nonpositive. The differences here are also significant for all nine products. Thus some ammount of historical smoothing appears to yield better estimates for every product in a significant majority of cells.

The last column of Table 1 reports which smoothing parameter appears most likely according to this simple count test. The parameter .3 is deemed best if $MS2_k-MS3_k$ is postive more often than nonpostive, but $MS3_k-MS4_k$ is not.

What is the gain from exponential smoothing? In equation (14) we see that the additional model mean squared error due to nonresponse is a multiple of σ_2^2. For cell k, $(1-(MSm_k/ML_k)) \times 100\%$ is a measure of the gain from exponentially smoothing February's reported value with a parameter of .m. It is literally the percent reduction in the nonresponse component of model mean squared error derived from this smoothing.

The test statistics for each method of determining historical values were aggregated across the cells so that the average gains could be displayed in a concise form. For example, the ML_k were aggregated to

$$ML=\frac{\sum_k(n_k-1)ML_k}{\sum_k(n_k-1)},$$

where n_k is the number of respondent units in cell k (after editing). It is interesting to note that ML is an estimator for the appropriate under the rather heroic assumption that $\sigma_2^2=\sigma_{2k}^2$ for all cells.

The measures of average gain relative to $\hat{X}_{Lk}$ for the nine products are displayed in Table 2. Negatives reflect average losses (increases in model mean squared error) rather than gains. The losses from using respondent mean imputation are also displayed.

After reviewing the two tables, one may conclude that for the retail categories using a smoothing parameter .4 would not be imprudent. This parameter offers gains of roughly 20% relative to updated historical imputation without smoothing.

For wholesale product, a smoothing parameter of .2 is better. The average gains are small, trivial for unleaded. A parameter of .3 works slightly better for premium, but the gain is still only 7.2%.

For all products, the best average gain tends to suggest a slightly higher smoothing parameter than the count test. This may be because the best value for λ varies from cell to cell with a median slightly smaller than its mean.

The results in the two tables do not appear to be sensitive to the month studied or to the size of the n_k. When only cells with $n_k>6$ were analyzed (roughly halving their number), the results were not qualitatively affected. Nor did an analysis based on August 1984 data yield significantly different results.

Table 1. Cell Counts

Product		Number of cells in which ME_k-ML_k is positive (nonpositive)	Number of cells in which ML_k-MSl_k is positive (nonpositive)	Which is the most likely smoothing parameter?
Sales Through Company Owned Outlets	Leaded	268 (18)	175 (111)	.4
	Unleaded	269 (13)	177 (105)	.3
	Premium	187 (20)	125 (82)	.4
Sales to Other End Users	Leaded	239 (38)	193 (84)	.4
	Unleaded	207 (43)	173 (77)	.4
	Premium	100 (23)	85 (38)	.3
Sales for Resale	Leaded	311 (31)	192 (150)	.2
	Unleaded	306 (27)	194 (139)	.2
	Premium	233 (24)	168 (89)	.4

Table 2. What is the Gain from Smoothing?

		Average Percent Reduction of Increase in Model MSE Due to Nonresponse Relative to $\hat{X}_{Lk}$					
Product		$\hat{X}_{S1k}$	$\hat{X}_{S2k}$	$\hat{X}_{S3k}$	$\hat{X}_{S4k}$	$\hat{X}_{S5k}$	$\hat{X}_{Ek}$
Sales Through Company Owned Outlets	Leaded	7.9	13.9	17.8	19.6	18.6	-2519
	Unleaded	8.8	15.5	20.0	22.3	22.2	-1683
	Premium	8.6	15.6	20.9	24.6	26.6	-1057
Sales to Other End Users	Leaded	8.3	14.7	19.2	22.0	22.8	-453
	Unleaded	8.1	14.4	19.2	22.5	24.1	-435
	Premium	7.2	12.9	16.9	19.2	19.4	-398
Sales for Resale	Leaded	2.3	3.8	4.4	4.0	2.4	-536
	Unleaded	0.5	0.1	-1.3	-3.7	-7.6	-585
	Premium	3.8	6.2	7.2	6.6	4.2	-738

CASSEL, C. M., SARNDAL C. E., and WRETMAN, J. H. (1977), Foundations of Inference in Survey Sampling, New York: John Wiley and Sons.
CASSEL, C. M., SARNDAL C. E., and WRETMAN, J. H. (1983), "Some Uses of Statistical Models in Connection with the Nonresponse Problem," in Incomplete Data in Sample Surveys, Volume 3: Proceedings of the Symposium, W. G. Madow and I. Olkin eds., 143-60.
FULLER, W. A. (1976), Introduction to Statistical Time Series, New York: John Wiley and Sons.
HUANG, E. T. (1985), "An Imputation Study for the Monthly Retail Trade Survey," American Statistical Association 1984 Proceedings of the Section on Survey Research Methods, 610-5.
ISAKI, C. T. and FULLER, W. A. (1982), "Survey Design Under the Regression Superpopulation Model," Journal of the American Statistical Association, 77, 89-96.
OH, H. L. and SCHEUREN, F. J. (1983), "Weighting Adjustment for Unit Nonresponse," in Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies, W. G. Madow, I. Olkin, and D. B. Rubin eds., 143-83.
The Petroleum Marketing Monthly, Washington: Energy Information Administration (any issue).

A mathematical appendix with proofs of the lemma and theorems contained herein is available from the author upon request.