BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-87/04

REPORT ON THE EFFECTS OF THE VIOLATION OF ASSUMPTION ON REGRESSION ESTIMATION OF CENSUS COVERAGE ERROR

by

Cary T. Isaki, Linda K. Schultz
Statistical Research Division
Bureau of the Census
Room 3134, F.O.B. #4
Washington, D.C. 20233 U.S.A.

Recommended: Kirk M. Wolter

Report completed: January 29, 1987

Report issued: January 29, 1987

Report on the Effects of the Violation of Assumptions on Regression Estimates of Census Coverage Error

by

C. Isaki, L. Schultz

I. Introduction

This report takes the regression model and formulation of the basic undercount model presented by Freedman and Navidi (1984) as a starting point. The authors specify seven assumptions underlying the undercount estimation procedure of Ericksen and Kadane (1985). They then discuss five potential violations of the assumptions. These violations deal with the synthetic assumption, bias of the direct estimates, omitted explanatory variables, measurement error of explanatory variables and estimation of variances (independence of errors).

Both of the above papers are concerned with regression modelling of percent net undercount of the 1980 Census using PEP (post enumeration program) state and large city estimates. We have at our disposal several artificial population counts defined by age-race-sex at the enumeration district level. The artificial populations were constructed by using the variable census substitutions (ratio adjusted so that the true population counts are equal to Demographic Analysis counts) as a proxy for undercount. A detailed description of the artificial populations is presented in Isaki, et.al. (1986). The main advantage in using the artificial population data over the PEP data is that the actual undercount as constructed is available with which to assess the effects of the violation of assumptions. The main disadvantage is that the proxy undercount variable may not sufficiently resemble undercount or lead to 1980 conditions affecting PEP estimation and subsequent modelling. This latter disadvantage may not be too severe because PEP estimation in 1990 is likely to differ from that done in 1980.

Consequently, we decided to use the artificial population data sets in our analysis. The artificial population data were used to assess the accuracy of the Demographic Analysis Synthetic and Statistical Synthetic estimators in previous work (Isaki, et.al. (1986), Isaki (1986a), Schultz, et.al. (1986)). Hence, by using the artificial populations in the present context we can also compare the regression related estimation results with the other two methods. Several points concerning our use of the artificial populations require comment. In our study of Statistical Synthetic estimation methods we developed a crude sample design to support estimation of the required adjustment factors. This allowed for obtaining the sampling covariance matrix of the estimated factors and also the selection of a replicate for exposition purposes. The sample design consisted of about 1400 enumeration districts (ED's), roughly 1.1 million persons, and provided estimates of 96 adjustment factors. These adjustment factors are presumed to be directly estimated from a PES. Hence, in the 1980 context and assuming use of Statistical Synthetic methodology, it is this set of 96 estimates that would have been modelled in a regression. Ideally, we would have preferred to investigate the violation of assumptions using the adjustment factors, however, we could not do so because we did not have an adequate set of explanatory variables at the adjustment factor domain level. Instead, we modelled state net undercount estimates of total population as measured by Statistical Synthetic 2 (syn 2) under the previously mentioned sample design. Syn 2 is defined in Isaki, et.al. (1986).

The advantages in using Syn 2 results in conjunction with our present aims are that explanatory variables for regression at the state level and a proxy for sampling bias are available. Also, a replicate estimate is available for use in comparing implementation results. In summary, all of the following results concerning violation of assumptions on regression estimates

are based on 1) artificial populations, 2) Syn 2 derived state estimates and 3) a given sample design. In the next section we display the seven assumptions as given by Freedman and Navidi (1984). We then discuss each of the violations of assumptions to be addressed in succeeding sections.

II. Model Assumptions and Notation

In the following, regression models of state percent net undercount of total population are developed. Net undercount is defined to be the ratio of the difference between the "measured" count and the census, and, the "measured" count. Thus, for state i,

$$Y_i = (T_i - C_i) / T_i \qquad 1)$$

$$Y_i' = (E_i - C_i) / E_i \qquad 2)$$

$$\hat{Y}_i = (\hat{E}_i - C_i) / \hat{E}_i \qquad \text{where} \qquad 3)$$

T_i = true population of state i as given by the artificial populations,

C_i = census population count of state i

E_i = Syn 2 estimate of total population for state i assuming adjustment factors are known without sampling error.

$\hat{E}_i$ = Syn 2 estimate of total population for state i assuming adjustment factors are measured with sampling error.

We thus consider three "measured" counts (C_i, E_i and $\hat{E}_i$) and hence we consider three different net undercounts. The first, Y_i, is the true net undercount. The second, Y_i', is the net undercount measured by the estimator E and is the expected value of $\hat{Y}_i$ apart from ratio bias. The third, $\hat{Y}_i$, is the sample based estimate of net undercount. In practice, one observes $\hat{Y}_i$ only. In the 1980 PEP, $\hat{Y}_i$ was constructed using the PEP state and sub-state estimates in place of $\hat{E}_i$ in 3). For the present we use the statisticsl synthetic 2 estimator of state total population as our $\hat{E}_i$. We now list the seven model assumptions as stated in Freedman and Navidi (1984) -

$$y_i = \gamma_i + \delta_i \qquad a)$$

$$\gamma_i = a + b\ min_i + c\ crime_i + d\ conv_i + \varepsilon_i \qquad b)$$

$$E[\delta_i] = 0 = E[\varepsilon_i] \qquad c)$$

$$Var\ \delta_i = K_i \qquad d)$$

$$Var\ \varepsilon_i = \sigma^2 \qquad e)$$

$$\delta_1, \delta_2, \ldots, \delta_{66}, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{66} \text{ are independent} \qquad f)$$

$$\delta_i \text{ and } \varepsilon_i \text{ are normally distributed.} \qquad g)$$

First observe that their discussion considers a set of 66 data points (states, balance of states, cities) while we will be concerned with 51 data points (states). Second, the y_i in a) is analogous to our $\hat{Y}_i$ if $\hat{E}_i$ is unbiased for T_i. Third, γ_i in b) is our Y_i and finally the explanatory variables in b), percent minority, crimes per thousand and percent conventional will be

We turn to a discussion of the assumptions a) to g) and how they relate to the artificial populations (AP2, AP3) and the synthetic estimates that we use in this report. In assumption a), $\hat{Y}_i$ is a biased estimator of Y_i with $E[\hat{Y}_i] \doteq Y_i'$. The range of this bias $(Y_i' - Y_i)$ is -1.88 to 1.50 for AP2 and -2.22 to 1.3 for AP3. The corresponding range for Y_i for AP2 are .19 to 7.71 and for AP3, .12 to 7.73. The bias of $\hat{Y}_i$ referred to in the above is the result of the weakness of the synthetic assumption underlying statistical synthetic 2 estimation. That is, that all persons in a designated group experience the same undercount rate. This is in contrast to other types of bias such as correlation and matching bias that apply in the case of dual system estimation. While such issues might apply in practice we assume for our present purposes that the dual system estimates that underly the statistical synthetic 2 estimator are unbiased. As for assumption b), we selected two explanatory variables each and fitted Y_i for both AP2 and AP3. The simple regressions were (all variables in percent)

$$Y_i^{AP2} = -1.65 + .369\ C5 + .0874\ C14\ ,\ \sigma^2 = .24 \qquad 4)$$

$$Y_i^{AP3} = -.72 + .0763\ C2 + .142\ C5\ ,\ \sigma^2 = .33 \qquad 5)$$

where

C5 is allocation as a percent of total population

C14 is minority renter as a percent of total population and

C2 is minority as a percent of total population

and minority is defined as Black or Hispanic.

We selected these explanatory variables because they seemed reasonable although other explanatory variables could have been used as well. Assumption c), that $E[\delta_i] = E[\varepsilon_i] = 0$ for every i does not hold for AP2 and AP3 for δ_i and is assumed to hold for the ε_i. Assumption d), known variance of the δ_i was also assumed to hold. In fact, we estimated the variance of $\hat{E}_i$ via replication and approximated the variance of δ_i via a first order Taylor's series expansion. An examination of the effect of assumption d) would have required a sample estimate of the variance of δ_i. Assumption e) was not examined. Rather, σ^2 was estimated via maximum likelihood under normality. A full examination of assumptions d) and e) could have been done (estimation of variances) but not within the given time frame. As for assumption f) we know that the $\hat{E}_i$ within census divisions are correlated and that between divisions they are not. We assumed independence of the ε_i. We did not use the covariances of the $\hat{Y}_i$ in our work. We assumed normality everywhere, assumption g). In summary, all results are conditional on the subset of assumptions of the a) to g) which we were not able to verify. The model we will work with and the assumptions are

$$\hat{Y}_i = Y_i' + \delta_i \qquad 6)$$

$$Y_i = \underset{\sim}{X}_i \underset{\sim}{\beta} + \varepsilon_i \text{ , } \underset{\sim}{X}_i \text{ described previously,} \qquad 7)$$

$$E[\delta_i] = E[\varepsilon_i] = 0 \text{ , } Var(\delta_i) \text{ is known , } Var(\varepsilon_i) = \sigma^2$$
is unknown and is to be estimated. 8)

$\delta_1, \ldots, \delta_{51}$, $\varepsilon_1, \ldots, \varepsilon_{51}$ are not independent (but we assume they are independent) and 9)

the δ_i and ε_i are normally distributed. 10)

Of the many possible violations of regression model assumptions we have addressed three of the ones specified by Freedman and Navidi (1984). The three that we will investigate are the effects of synthetic application of regression, effects of bias in the $\hat{Y}_i$ and the effects of excluding an explanatory variable in the regression model.

III. Synthetic Application of Regression

Synthetic application of regression refers to the use of a regression developed at the state level to predict at a different level, such as counties where direct estimates are not felt to be reliable. To examine this issue (for both AP2 and AP3) we used a weighted regression model of $\hat{Y}_i = \underset{\sim}{X}_i \underset{\sim}{\beta} + \eta_i$ where $\eta_i \sim N(0, \sigma^2\underset{\sim}{I} + \underset{\sim}{D})$ $i = 1, 2, \ldots 51$ and $\underset{\sim}{D}$ is a diagonal matrix with sampling variances on the diagonal. The resulting regression estimates were estimated to be

$$\hat{Y}_i^{AP2} = -.7086 + .2241\ C5 + .0957\ C14\ ,\ \hat{\sigma}^2 = .083 \qquad 11)$$

$$\hat{Y}_i^{AP3} = -.2569 + .0691\ C2 + .0939\ C5\ ,\ \hat{\sigma}^2 = .003\ . \qquad 12)$$

The next step was to predict net undercount for each county using 11) and 12). Finally, to obtain county predicted total population, net undercount was "unraveled". That is, net undercount was converted to a total population figure using the census count. In the E/K methodology, the directly estimated net undercount figure for a state, say, is averaged with the predicted net undercount. Hence, an E/K estimate of total refers to a conversion (unraveling) of this averaged figure. Also, the elements of $\underset{\sim}{D}$ above, the sampling variances of the directly estimated net undercount figures were approximated using a first order Taylor's series expansion and the variance of the directly estimated total population state figure. Because the $\hat{Y}_i$ are net

undercount estimates of state i using the Syn 2 estimator, the dependent variables in 11) and 12) are biased estimates of Y_i, the actual net undercount. Despite this additional complication, the results remained favorable toward synthetic application of regression.

For comparison purposes, the county total population estimates were input into several measures of improvement used previously. In addition to direct use of 11) and 12), two other alternatives were considered. The alternatives use ratio adjustments by state. Finally, a simple ratio adjustment of census counts is also considered.

The alternatives are

$$\hat{Y}_{C_{1i}} = \left[\hat{Y}_i^{AP} \Big/ \sum_{i \varepsilon s} \hat{Y}_i^{AP}\right] \hat{Y}_s^{E/K} \quad \text{and} \qquad 13)$$

$$\hat{Y}_{C_{2i}} = \left[\hat{Y}_i^{AP} \Big/ \sum_{i \varepsilon s} \hat{Y}_i^{AP}\right] \hat{Y}_s^{dse} \qquad 14)$$

where

$\hat{Y}_i^{AP}$ is the predicted county total population for county i

$\hat{Y}_s^{dse}$ is the Syn 2 state total population estimate for state s

$\hat{Y}_s^{E/K}$ is the predicted state total population using E/K methodology.

The simple ratio adjustment of census counts estimator is

$$\hat{Y}_{C_{3i}} = \left[c_i \Big/ \sum_{i \varepsilon s} c_i\right] \hat{Y}_s^{dse} \qquad 15)$$

while for completness, the non-ratio adjusted regression estimator is

$$\hat{Y}_{C_{0i}} = \hat{Y}_i^{AP} \quad . \qquad 16)$$

At the risk of some confusion we continue to use the subscript i in lines 13) - 16) above. While prior to 13) i denoted state i, for the remainder of Section III we use s to denote state s and i to denote the i-th county. We briefly describe the four estimators in the above -

i) $\hat{Y}_{C_{0i}}$ is the predicted total population for county i obtained by first using the regression equation in 11), say, and the relevant explanatory values for the county and then converting net undercount to a level figure.

ii) $\hat{Y}_{C_{1i}}$ and $\hat{Y}_{C_{2i}}$ are both ratio estimators that ratio adjust $\hat{Y}_{C_{0i}}$ but to two different state totals. The first state total is one obtained using Ericksen and Kadane's empirical Bayes estimator while the second state total is that obtained by using the Syn 2 estimator of total.

iii) $\hat{Y}_{C_{3i}}$ is a ratio estimator that ratio adjusts the census count for county i in state s to the Syn 2 state estimate of total population.

The results in Table 1 and Table 2 provide an illustration of the results of an application of adjustment methods at the county level. Among the methods considered, either C_0, no ratio adjustment of the regression, or C_1, using the E/K estimator for ratio adjustment of the regression are preferable to C_2, C_3 or the census. The results for C_0 and C_1 are similar because from 11) and 12), the magnitude of $\hat{\sigma}^2$ causes the E/K estimator to be

Table 1. Measures of Improvement Applied to County (3137 Counties) Adjusted Total Population Based on C_0, C_1, C_2, C_3 and the Census for Artificial Population 2 (AP2).*

	Measure	C_0	C_1	C_2	C_3	Census
1.	No. of counties where ARE(census) < ARE(C_i)	1040	1072	1208	1442	-
2.	No. of counties where ADP(census) < ADP(C_i)	1279	1275	1278	1545	-
3.	MARE	.0081	.0082	.0093	.0108	.0128
4.	Max ARE	.2074	.2078	.2114	.2179	.2236
5.	Median ARE	.0048	.0050	.0058	.0071	.0076
6.	α	33695	32430	38013	52101	115609
7.	RSADP	1.198	1.198	1.198	.993	-
8.	PI	.614	.614	.614	.517	-
9.	ϕ	33550	32304	37806	51895	55664
10.	MP1 $x10^{-2}$	.573	.572	.573	.771	.771

Table 2. Measures of Improvement Applied to County (3137 counties) Adjusted Total Population Based on C_0, C_1, C_2, C_3 and the Census for Artificial Population 3 (AP3).*

	Measure	C_0	C_1	C_2	C_3	Census
1.	No. of counties where ARE(census) < ARE(C_i)	1141	1115	1286	1626	-
2.	No. of counties where ADP(census) < ADP(C_i)	1152	1149	1154	1534	-
3.	MARE	.0075	.0075	.0086	.0106	.0111
4.	Max ARE	.2673	.2673	.2828	.3003	.3069
5.	Median ARE	.0035	.0034	.0049	.0064	.0055
6.	α	37204	37206	48639	73514	134494
7.	RSADP	1.289	1.289	1.289	.9997	-
8.	PI	.682	.682	.682	.5088	-
9.	ϕ	37163	37179	48161	73036	74199
10.	MP1 $x10^{-2}$	.691	.691	.691	1.007	1.007

*For a definition of the measures, see the Appendix.

the regression essentially and to the extent that additivity held, the ratio factor would be close to one. At any rate, C_2 performed better than C_3 and both performed better than the census for both AP2 and AP3. Synthetic application of regression has been shown to be superior to the census for AP2 and AP3. A comparison between the performance of C_0 and syn 2 and syn DA (see Isaki, et.al. 1986) under the same conditions (replicate) and for county estimation is favorable toward C_0. This further supports the case that for AP2 and AP3 a synthetic regression assumption application produces an improved adjustment figure (for counties). Further investigation would be required as to defining the small area level limits of such an assumption.

One way to examine the limits of the synthetic regression assumption is to examine the county level adjustments by size categories. We grouped the C_0 adjustment results, to illustrate the trend, into 3 county sizes - about 0 to 10,000; 10,001 to 50,000; 50,001+ persons with number of counties 784, 1569 and 784, respectively. The results for AP2 and AP3 are contained in Tables 3 and 4. The measures progressively favor adjustment as the size category increases. In fact, all measures favor adjustment over the census, even those for counties with population under 10,000. A trend is visible, however, and further investigation is warranted to approximate the small area level at which adjustment is not superior to the census.

Table 3. Measures of Improvement Applied to County Adjusted Total Population Based on Method C_0 and the Census for Artificial Population 2 and by Total Population Size Groups (0 to 10,000; 10,001 to 50,000; 50,001+) With Number of Counties (784; 1569; 784).

	Measure	0 to 10,000		10,001 to 50,000		50,001+	
		Census	C_0	Census	C_0	Census	C_0
1.	No. of counties where ARE(census) < ARE(C_i)	-	365	-	494	-	189
2.	No. of counties where ADP(census) < ADP(C_i)	-	258	-	504	-	205
3.	MARE	.0110	.0091	.0136	.0085	.0131	.0064
4.	Max ARE	.2236	.2072	.1834	.1467	.1281	.0879
5.	Median ARE	.0052	.0051	.0081	.0049	.0087	.0041
6.	α	2258	1247	18671	8363	94864	24069
7.	RSADP	-	1.239	-	1.293	-	1.506
8.	PI	-	.657	-	.685	-	.718
9.	ϕ	1584	1247	11616	8331	42101	23834
10.	MP1 $x10^{-3}$	.343	.263	.313	.220	.235	.129

Table 4. Measures of Improvement Applied to County Adjusted Total Population Based on Method C_0 and the Census for Artificial Population 3 and by Total Population Size Groups (0 to 10,000; 10,001 to 50,000; 50,001+) With Number of Counties (784; 1569; 784).

	Measure	0 to 10,000		10,001 to 50,000		50,001+	
		Census	C_0	Census	C_0	Census	C_0
1.	No. of counties where ARE(census) < ARE(C_i)	-	364	-	570	-	220
2.	No. of counties where ADP(census) < ADP(C_i)	-	218	-	498	-	162
3.	MARE	.0097	.0079	.0116	.0080	.0117	.0061
4.	Max ARE	.1814	.1732	.3069	.2672	.1253	.0730
5.	Median ARE	.0037	.0034	.0055	.0035	.0067	.0035
6.	α	1874	1019	19769	10355	112896	25794
7.	RSADP	-	1.366	-	1.379	-	1.701
8.	PI	-	.710	-	.685	-	.745
9.	ϕ	1372	1004	14592	10355	57232	25794
10.	MP1 $x10^{-3}$	.297	.211	.394	.272	.319	.139

IV. Effects of Bias in the Population Estimates

To examine the effects of bias in the state total population estimates the differences between percent net undercount using the E_i figures of state population in which the adjustment factors are known without error Y_i' and the true percent net undercount Y_i was examined. The difference between the terms, $Y_i' - Y_i$, termed the bias of net undercount was found to be related to the explanatory variables used to model the true net undercount Y_i. This implies that when the estimated percent net undercount is modelled the bias is also being modelled. Unlike Freedman and Navidi's discussion of bias we are able, because of the known artificial population, to examine the effect of bias on the measures defined in the appendix. In the following discussion we compare the modelled true undercount with the modelled syn 2 undercount after the estimates are "unraveled" to obtain state estimates of population. By examining the predicted undercount estimate of state population based on known adjustment factors rather than estimated factors, we eliminate one source of variation, namely the variation due to sample estimates. Comparing to the true undercount we are able to get a handle on the bias and its effects on the measures.

The simple linear models used to investigate this issue are described below. For AP2 and AP3 the models fit to the true undercount using the explanatory variables previously defined are described in equations 4) and 5) and are denoted $\hat{AP2}$ and $\hat{AP3}$ in Tables 5 and 6 below. We also fit the undercount estimates computed with known adjustment factors with the same explanatory variables

$$Y_i^{,AP2} = -.46 + .204\ C5 + .088\ C14 \quad , \quad \sigma^2 = .381 \qquad 17)$$

$$Y_i^{,AP3} = -.0884 + .0586\ C2 + .142\ C5 \quad , \quad \sigma^2 = .313 \qquad 18)$$

The results of this application are denoted $\hat{EE}$ in Tables 5 and 6 below.

Table 5. Measures of Improvement Applied to State Adjusted Total Population Based on Simple Linear Regression Model for the True Undercount (AP2) and for the Undercount Based on the Syn 2 Model with Known Adjustment Factors (EE) for Artificial Population 2 (AP2)*

	Measure	$\hat{AP2}$	$\hat{EE}$	Census
1.	No. of states where ARE(census) < ARE(adj)	1	5	-
2.	No. of states where ADP(census) < ADP(adj)	0	0	-
3.	MARE	.0037	.0042	.0147
4.	Max MARE	.0103	.0197	.0771
5.	Median ARE	.0027	.0029	.0113
6.	α	7803	8007	77316
7.	RSADP	1.432	1.515	-
8.	PI	.636	.698	-
9.	ϕ	7548	7905	17391
10.	MP1 $x10^{-3}$	.0332	.0346	.0788

Table 6. Measures of Improvement Applied to State Adjusted Total Population Based on Simple Linear Regression Model for the True Undercount (AP3) and for the Undercount Based on the Syn 2 Model with Known Adjustment Factors (EE) for Artificial Population 3 (AP3)*

	Measure	$\hat{AP3}$	$\hat{EE}$	Census
1.	No. of states where ARE(census) < ARE(adj)	3	7	-
2.	No. of states where ADP(census) < ADP(adj)	12	11	-
3.	MARE	.0040	.0046	.0136
4.	Max MARE	.0173	.0240	.0773
5.	Median ARE	.0023	.0034	.0092
6.	α	7803	7956	82365
7.	RSADP	1.828	1.820	-
8.	PI	.641	.655	-
9.	ϕ	7752	7854	22032
10.	MP1 $x10^{-3}$	.034	.035	.100

*For a definition of the measures, see the Appendix.

The results in Table 5 and Table 6 are based on simple linear regression models in which the explanatory variables were chosen based on their relationship to the true undercounts respectively for AP2 and AP3. For each artificial population, true undercount was fit to the appropriate explanatory variables as was undercount produced from syn 2 estimates formed with adjustment factors without error (Y_i'). This allowed the effect of bias on the measures of improvement to be examined.

Tables 5 and 6 indicate that the bias in the syn 2 known factor estimates of state population does not have a substantial effect on the measures of improvement for AP2 and AP3. Furthermore the reader is reminded that the range of bias is (-1.88, 1.50) percent for AP2 and (-2.22, 1.3) for AP3 while the actual net undercount in percent were (.19, 7.71) and (.12, 7.73) for AP2 and AP3 respectively over all states. Under the conditions examined, we can say that the bias due to statistical synthetic estimator 2 that arises due to failure of the synthetic assumption (all units in the same adjustment strata have the same undercount rate) has not degraded the adjustment reaults. All measures indicate that any set of adjusted state counts described here is superior to the census. Finally, it is emphasized that the bias considered here does not include bias arising from failure of the assumptions underlying dual system estimation.

V. Effects of Excluding an Explanatory Variable in the Regression Model

To examine the effects of excluding an explanatory variable from the regression model one of the two explanatory variables from equations 11) and 12) were dropped. For both artificial populations, variable C5 was the variable selected to be dropped. As with equations 11) and 12) the reduced models were fit with a weighted regression $\hat{Y}_i = \underset{\sim}{X}_i \underset{\sim}{\beta} + \eta_i$ where

$\eta_i \sim N(0, \sigma^2 \underset{\sim}{I} + \underset{\sim}{D})$. The resulting regression estimates were

$$\hat{Y}_i^{AP2} = .539 + .141\ C14, \quad \sigma^2 = .1639 \qquad 19)$$

$$\hat{Y}_i^{AP3} = .267 + .078\ C2, \quad \sigma^2 = .0119\ . \qquad 20)$$

To compare the results of equations 11) and 19) for AP2 and the results of equations 12) and 20) for AP3 the predicted state percent net undercount estimates were converted using the appropriate census counts to obtain state population estimates. The state population estimates were then compared, using the measures of improvement defined in the appendix, to examine the effect of not including an explanatory variable that should have been included in the model. Examination of the effect of missing an explanatory variable in regression was done by way of comparisons in the presence of sampling error. An alternative would have been to exclude sampling error. The former situation is expected to apply. The results for both AP2 and AP3 are provided in the tables below.

Table 7. Measures of Improvement Applied to State Adjusted Total Population to Examine the Effects of Excluding An Explanatory Variable for Artificial Population 2 (AP2)*

	Measure	eq. 11) 2 Variables	eq. 19) 1 Variable	Census
1.	No. of states where ARE(census) < ARE(adj)	5	6	-
2.	No. of states where ADP(census < ADP(adj)	11	11	-
3.	MARE	.0039	.0047	.0147
4.	Max ARE	.0131	.0183	.0771
5.	Median ARE	.0025	.0041	.0113
6.	α	8160	12546	77316
7.	RSADP	1.479	1.223	-
8.	PI	.646	.685	-
9.	ϕ	8007	12240	17391
10.	MP1 $x10^{-3}$	.0351	.0537	.0788

*For a definition of the measures, see the Appendix.

Table 8. Measures of Improvement Applied to State Adjusted Total Population to Examine the Effects of Excluding An Explanatory Variable for Artificial Population 3 (AP3)*

	Measure	eq. 12) 2 Variables	eq. 20) 1 Variable	Census
1.	No. of states where ARE(census) < ARE(adj)	7	7	-
2.	No. of states where ADP(census) < ADP(adj)	12	11	-
3.	MARE	.0041	.0043	.0136
4.	Max ARE	.0186	.0194	.0773
5.	Median ARE	.0023	.0031	.0092
6.	α	7650	8262	82365
7.	RSADP	1.862	1.800	-
8.	PI	.639	.672	-
9.	ϕ	7650	8211	22032
10.	MP1 $x10^{-3}$	.0335	.0361	.1000

*For a definition of the measures, see the Appendix.

The results in Table 7 and Table 8 indicate that for all but one of measures of improvement examined that the ability to predict population accurately is diminished by dropping an explanatory variable from the mo The results also show that for both AP2 and AP3 that both the one and th variable models are superior to the census.

The issue of measurement error of explanatory variables in regresssi and its effects on adjustment was mentioned by Freedman and Navidi as wel In our application to date and recognizing the tight adjustment schedule, have litittle hope that long form questionnaire items (subject to samplin error) will be available for modelling. Assuming that short form items w have minimal measurement error, we have omitted consideration of this pr for now.

APPENDIX

Measures of Improvement

Let

c denote the census population count

s denote the true population count

e denote the estimated population count.

1. Number of areas where ARE(c) < ARE(e)

   where

   $$ARE(c) = |(c-s)/s|$$

   $$ARE(e) = |(e-s)/s|$$

2. Number of areas where ADP(c) < ADP(e)

   where

   $$ADP(c) = |P^c - P^s|$$

   $$ADP(e) = |P^e - P^s|$$

   $$P^c = {}^{c_i}/_{\Sigma^N c_i} \text{ for the i-th area}$$

   $$P^s = {}^{s_i}/_{\Sigma^N s_i} \quad , \quad P^e = {}^{e_i}/_{\Sigma^N e_i}$$

3. $$MARE = \frac{1}{N} \sum_{i}^{N} \left| \frac{e_i - s_i}{s_i} \right|$$

4. Maximum ARE(e) = Max ARE(e)

5. Median ARE(e)

6. Weighted squared relative error

$$\alpha = \sum_{i}^{N} s_i \left[(e_i - s_i) / s_i\right]^2$$

7. $$\text{RSADP} = \sum_{i}^{N} \left| P_i^{c} - P_i^{s} \right| / \sum_{i}^{N} \left| P_i^{e} - P_i^{s} \right|$$

where $P_i^{c} = \dfrac{c_i}{\sum_{i}^{N} c_i}$, etc.

8. $$\text{PI} = \sum_{i}^{N} \text{IMPV}_i / M$$

$$M = \sum_{i}^{N} s_i \qquad \text{IMPV}_i = \begin{cases} s_i & \text{if } \left|P_i^{e} - P_i^{s}\right| < \left|P_i^{c} - P_i^{s}\right| \\ 0 & \text{otherwise} \end{cases}$$

9. Weighted squared relative error differences

$$\phi = \sum_{i}^{N} s_i \left[\left\{(e_i - s_i)/s_i\right\} - \left\{\left(\sum_{i}^{N} e_i - \sum_{i}^{N} s_i\right)/\sum_{i}^{N} s_i\right\}\right]^2$$

10. $$\text{MP1} = \sum_{i=1}^{N} \frac{(P_i^{s} - P_i^{e})^2}{P_i^{s}}$$

VI. References

1. Ericksen, E.P. and Kadane, J.B. (1985), "Estimating the Population in a Census Year - 1980 and Beyond," Journal of the American Statistical Association, 80, 98-109.

2. Freedman, D.A. and Navidi, W.C. (1984), "Regression Models for Adjusting the 1980 Census," Technical Report No. 35, University of California, Berkeley, Dept. of Statistics.

3. Isaki, C.T., Diffendal, G.J. and Schultz, L.K. (1986), "Statistical Synthetic Estimates of Undercount for Small Areas", Proceedings of the Bureau of the Census' Second Annual Research Conference, pg. 557-569, Reston, Virginia.

4. Isaki, C.T. (1986a), "Report on Statistical Synthetic Estimation", draft of internal report, Statistical Research Division, Bureau of the Census.

5. Schultz, L.K., Huang, E.H., Diffendal, G.J. and Isaki, C.T. (1986), "Some Effects of Statistical Synthetic Estimation on Census Undercount of Small Areas," paper presented at the Annual Meeting of the American Statistical Association, Survey Methodology Section, Chicago, Illinois.