DOWNWEIGHTING INFLUENTIAL CLUSTERS IN SURVEYS, WITH APPLICATION TO THE 1990 POST-ENUMERATION SURVEY

Thomas R. Belin, Departments of Psychiatry and Biostatistics, UCLA,
Nathaniel Schenker, National Center for Health Statistics and Alan M. Zaslavsky, Harvard Medical School
Alan M. Zaslavsky, Department of Health Care Policy, 180 Longwood Ave., Boston, MA 02115-5899*

Key words: Census undercount, infinitesimal jackknife, influence, outliers, robustness, t distribution

Abstract

Certain clusters may be extremely influential on survey estimates from clustered samples and consequently contribute disproportionately to their variance. We propose a general approach to downweighting clusters using a robust estimation strategy based on M-estimation, using t-based weight functions. The method is motivated by a problem in census coverage estimation. In this context, both extreme weights and large errors can lead to extreme influence, and influence can be estimated by Taylor linearization. As predicted by theory, the robust procedure greatly reduces the variance of estimated coverage rates, more so than truncation of weights. On the other hand, the procedure may introduce bias into survey estimates when the distributions of the influence statistics are asymmetric. We demonstrate techniques for assessing the bias-variance tradeoff and consider the properties of the estimators in the presence of asymmetry. We also suggest design improvements to reduce the impact of influential clusters.

1 Introduction

In clustered samples, certain clusters may be extremely influential on a survey estimate and consequently contribute disproportionately to its variance. As noted in the review by Lee (1995), a cluster may be influential because it has an extreme sampling or poststratification weight compared to the weights for other clusters in the same area or containing similar population groups. A cluster may also be influential because it is an outlier, i.e., because some measured quantity of interest is extreme relative to a postulated distribution for that quantity across clusters. Chambers (1986) distinguishes between extreme values that are incorrectly recorded ("nonrepresentative outliers") and those that are correctly recorded ("representative outliers"). Here we presume that incorrectly recorded values have been edited in the data being analyzed.

Previous research on controlling influential observations in surveys can be classified into two groups: research on methods for handling outlying data values; and research on methods for handling extreme weights. We summarize briefly here; BSZ gives references. With regard to handling outliers, one strategy is to identify and edit them, possibly using the influence function as a diagnostic. A second strategy for dealing with outliers is to apply robust estimation techniques to the data. A common strategy for handling extreme weights is weight trimming (Potter 1990), which involves identifying a ceiling on allowable weights.

In this paper, we develop an approach to downweighting clusters based on techniques from the theory of robust estimation, also related to jackknife estimation. For a given estimator, a derivative-based influence statistic is calculated for each cluster, which represents the amount by which the estimator would change if the cluster were dropped from the data. A new estimator is then calculated using modified weights, where the modification factors are determined by fitting a t distribution to the influence statistics.

Our approach differs from the robust estimation approaches cited above, in that the estimation technique is applied to the influence statistics rather than to the raw data values. This provides a unified treatment of extreme weights and outliers, which is desirable since they determine influence together. For example, a cluster with a severely outlying data value but a somewhat low weight might only have moderately high influence. A robust method that is applied directly to the raw data values might downweight such a cluster severely, and weight trimming might not downweight the cluster at all. In contrast, our method would downweight the cluster moderately, which seems appropriate given the level of influence of the cluster. Modeling the influence statistics directly also addresses influence due to more than one variable at the same time (e.g., census undercount and overcount in the application presented in this paper).

*Author listing is alphabetical. The authors thank Gregg J. Diffendal of the Bureau of the Census for developing the data set analyzed in this paper and for his immense contribution to our understanding of the 1990 PES. The authors also thank Robert E. Fay III, Roderick J.A. Little, and Hyunshik Lee for comments that helped to improve the paper. The proceedings version of this paper has been shortened and most tables and figures have been removed; a more complete version (referred to as BSZ in the text) is available by request to zaslavsk@hcp.med.harvard.edu.

Our research is motivated by the problem of estimating coverage in the decennial census. We illustrate our techniques using cluster-level data from the 1990 Post-Enumeration Survey (PES). As predicted by theory, the t-based procedure leads to large reductions in variance, but it may introduce bias into survey estimates due to asymmetry of the distribution of influence statistics. We demonstrate techniques for assessing the bias-variance tradeoff and consider the properties of the estimators when the underlying distributions are asymmetric.

Section 2 discusses the 1990 PES, how the PES was used to estimate coverage in the census, and sources of influential clusters in the PES. Section 3 presents a general formulation of our approach to downweighting influential clusters. The approach is applied to data from the 1990 PES in Section 4. We conclude by discussing areas for further research (Section 5).

2 Influential Clusters in the 1990 Post-Enumeration Survey

2.1 Overview

Coverage error in the 1990 United States Census of Housing and Population was estimated using a post-enumeration survey (PES), a stratified cluster sample in which the primary sampling units were census blocks (typically either city blocks or rural areas containing several housing units) or groups of census blocks (Hogan 1993). The design and the processing of the PES caused some clusters to be very influential in the estimation of coverage error. One set of influential clusters were those where large-scale errors in the census were detected by the PES. Among these were clusters in which unusually many households were misgeocoded (assigned to the wrong geographic location) or missed altogether. Other clusters were influential because they had very high sampling weights. We postpone a formal definition of influence to Section 3, but we note that in the context of census coverage estimation, the influence of a cluster is roughly proportional to the excess of the weighted number of cases contributed by the block to the estimated total undercount over the number that would be expected at the general undercount rate for the poststratum.

Some of the processes that result in influential clusters would be expected to yield equally many large contributors to estimated undercount and overcount. Other such processes would not yield such a balance. In either case, influential clusters can contribute disproportionately to the variance of estimates of coverage error. In estimating coverage for the 1990 census, the Census Bureau reduced the influence of certain PES clusters by truncating their weights on a post hoc basis. In Section 4, we explore our alternative t-based approach and compare it to some simple schemes for truncating weights.

2.2 Coverage Estimation Methodology

The 1990 PES consisted of two parts: a sample of the population called the P sample and a sample of census enumerations called the E sample. The P sample was used to estimate the proportion of the population that was missed in the census, whereas the E sample was used to estimate the number of erroneous enumerations in the census.

The 1990 PES was a stratified sample of 5,290 block clusters. The P sample consisted of all people who lived in the sample clusters at the time of the PES interview and should have been counted in the census. The E sample consisted of all enumerations that the census placed in the same sample clusters.

Clusters were sampled with known probabilities, with sampling weights equal to inverse probabilities of selection. In general, the weight for a cluster was applied to all the individuals in the cluster, although weighting adjustments were performed for households where no interview was obtained (Belin, Diffendal, Mack, Rubin, Schafer, and Zaslavsky 1993). In certain clusters with large populations, subsampling was carried out to reduce field work, and weights were modified accordingly so that perhaps half as many households would be interviewed but their weights would be doubled.

Special consideration was also given in the sample design to "small blocks," defined from a pre-census housing unit count for every census block in the country. Small blocks include business areas, median strips of highways, parks, rural areas, and bodies of water where people might dwell. The original plan for the 1990 PES included two sample small blocks for each of about 50 strata defined by geography, but concerns about effects on variances of the large weights of small blocks led to the inclusion of a supplemental sample of about 3,000 small blocks. This supplemental sample was listed close to the time of PES interviewing, and block clusters with ten or more housing units in either the P or E sample were included in the PES.

To provide data for estimating the proportion of the population that was missed in the census, the PES determined where each P-sample person lived on the reference day of the census. The P sample was then matched against the census through a combination of computer and clerical matching operations. A P-sample person was considered a census enumeration if he/she had been enumerated in the census within a search area composed of the census block reported in the PES and a ring of surrounding census blocks (two rings in rural areas). Individuals found in the P sample but not in the census ("nonmatches") were followed up to confirm their existence.

Erroneous enumerations in the census included duplicates, fictitious individuals, persons not alive at the time of the census, and persons counted in the wrong location ("geocoding errors"). Enumerations in the E sample were checked against the census to determine whether they were duplicates. In addition, E-sample cases that did not have matches in the P sample were followed up to determine whether they were erroneous enumerations other than duplicates.

E-sample geocoding errors were defined by a rule similar to that used for P-sample cases. With this rule, it was equally likely that a housing unit misgeocoded in the census outside the search area for its correct location would appear as an erroneous enumeration (if the erroneous location was in a sample cluster) or as a census omission (if the true location was in a sample cluster). Thus, the E- and P-sample rules "balanced" each other. Of the 21,063 cases whose geocoding status was determined from follow-up operations, 42% were correctly geocoded in the E-sample block, 52% were classified as belonging to a census block adjacent to the E- sample census block and thus were correctly geocoded in the search area, and 6% were erroneously geocoded.

Estimation poststrata were defined by geography, race/ethnicity, tenure (i.e., owner/renter status), age, and sex. A sample cluster would typically fall into one geographic area but contain persons in several poststrata. The estimator of the adjustment factor is given below by (2). Details of the design of the PES and the 1990 coverage estimation methodology appear in Hogan (1993).

2.3 Sources of Influential Clusters

In this section, we briefly describe some sources of influential clusters in the 1990 PES. For a more detailed discussion and a descriptive analysis of these clusters, see Diffendal, Zaslavsky, Belin, and Schenker (1994) or BSZ.

The clustered design of the 1990 PES facilitated field operations and matching, but also permitted cluster-level errors to affect the accuracy of the survey. The results of the 1990 PES included over three dozen clusters in which there was a particularly poor match between census and PES rosters. These clusters were outliers in relation to general patterns of error. In other words, the high levels of nonmatch were not due simply to a generally high rate of census and PES errors in the area at the person or household level, but rather to specific large-scale errors that affected whole clusters or substantial portions of them.

Some of these large-scale errors were due to problems in field operations. For example, a substantial portion of a cluster could have been missed by the census. Other large-scale errors were due to problems in geocoding. For example, an entire housing development or apartment building in the P sample might have been geocoded outside the corresponding search area in the census, causing all of its residents to be classified as nonmatches in the P sample. Conversely, a similar collection of households might have been geocoded erroneously into an E-sample block from outside the search area. Although geocoding errors should balance out in expectation, and although PES matching rules are designed specifically to enforce this balance across the PES sample (Section 2.2), particular poststrata may be greatly influenced by such errors since most of the population of each cluster falls into only a few poststrata.

Even in the absence of a large-scale error, a cluster could still be very influential because of extreme weights, defined as sampling weights that are very large compared to weights for other clusters in similar areas or with similar population groups. (Conversely, a large-scale error might not result in high influence if it occurs in a cluster with a very small weight.) For example, the intention of the 1990 PES sample design was that the high-weight "small blocks" should have little population. In fact, due to errors in precensus listing and the census itself, some had substantial counts. In combination with their large weights, this made them very influential. The errors in these blocks were not necessarily large in absolute terms, but they were large in relation to the anticipated populations of the blocks.

With a sample of a few thousand clusters, there is not enough information to identify accurately the systematic effects of large-scale errors and extreme weights, e.g., that they lead more frequently to urban residents being placed in suburban areas than the other way around. Therefore, downweighting influential clusters, as was done in 1990, is appropriate. Post-hoc procedures for downweighting are subject to criticism, however, because of their reliance on expert judgement applied to individual blocks and because of the discontinuity between those blocks selected for downweighting and those not selected. The approach outlined in Sections 3 and 4 attempts to provide an objective and systematic basis for downweighting.

3 Robust Estimation When There are Influential Clusters

We next present models and theoretical perspectives that suggest methods that reduce the contribution of influential clusters to variance and thereby potentially improve the quality of survey estimates. We first discuss how, in the context of variance estimation, estimates of the undercount (or other nonlinear multivariate statistics) may be approximated as a mean of values for each observation, the influence. We next review robust estimation of the mean using M-estimators. Finally, we propose combining these tools by using a robust estimator applied to influence statistics.

3.1 The Infinitesimal Influence

The "influence" of a unit (in our paper, a cluster) on an estimator may be defined as the negative of the amount by which the estimator would change if the unit were dropped from the data. Influence may be calculated directly by removing each unit from the data and recalculating the estimate, as in jackknife methods for variance estimation.

In some cases, an approximately equal influence measure may be obtained in closed form as a function of data values by taking the derivative of the estimator with respect to the inclusion indicator for each unit. Formally, suppose that the estimator can be written as $g(X, I)$, where $X = (x_1, \ldots, x_N)^T$ is the matrix of data values in the population with x_i being the data vector for unit i, and $I = (I_1, \ldots, I_N)^T$ is the vector of inclusion indicators ($I_i = 1$ if unit i is included in the sample, $I_i = 0$ otherwise). Then the derivative-based, or "infinitesimal," influence measure is defined for units included in the sample as

$$D_i = \frac{\partial g(X, I)}{\partial I_i}. \qquad (1)$$

Because of the close theoretical and practical correspondence between this measure and the jackknife, it was called the "infinitesimal jackknife" in early research (Church and Harris 1970; Jaeckel 1972). This statistic is the basis of "Taylor linearization" approaches to variance estimation.

Variance estimates based on the jackknife, which is rooted in survey sampling theory, and based on the infinitesimal influence, which is an important tool in the theory of robust estimation, are essentially equivalent if the estimator is a smooth function of the data (Jaeckel 1972; Efron 1982). In both cases, we may estimate variance as if we were calculating the sum of the influence statistics themselves rather than an arbitrary estimator based on the original data. Jackknife theory tells us that this equivalence holds with weighted sampling schemes as well (ignoring finite population corrections in without-replacement sampling schemes, which are negligible in many applications).

3.2 Long-Tailed Distributions and Robust Estimation

It is well known that the sample mean is the optimal estimator of location for a normally distributed population. At the opposite extreme, with very long-tailed distributions (specifically, the Laplace or double-exponential distribution), the optimal estimator is the median, which gives no weight to any observation other than the middle one.

A large class of robust estimators of location is the M-estimators (Huber 1964; Huber 1981; Hampel, Ronchetti, Rousseeuw, and Stahel 1986). For a location-scale family, these are defined by the optimization $\hat{\mu} = \arg\min_\mu \sum_i \rho((y_i - \mu)/\sigma)$, or equivalently by the estimating equation $\sum_i \psi((y_i - \mu)/\sigma) = 0$, where $\psi = \rho'$. If ρ is a loglikelihood for a distribution, the M-estimator is simply a maximum likelihood estimator, but the usefulness of the estimator does not depend on whether a specific distributional assumption holds.

The M-estimator of μ may be calculated as an iteratively weighted mean, $\hat{\mu} = \sum_i w_i y_i$, where $w_i = \psi((y_i - \mu)/\sigma)/(y_i - \mu)$ depends on the previous estimate of μ and σ. Robust M-estimators, such as those based on long-tailed distributions, give reduced weight to the extreme observations. This downweighting is the source of the robustness of the estimator against outliers. The corresponding estimator of scale is derived by differentiating $\sum_i \rho((y_i - \mu)/\sigma)/\sigma$ with respect to σ and equating the derivative to 0. Both $\hat{\mu}$ and $\hat{\sigma}$ are updated in each iteration.

One suitable family of long-tailed distributions for defining an M-estimator is the t family, because the degrees-of-freedom parameter ν allows us to approximate the tail shape of an observed distribution between the extremes of the Cauchy (very heavy tails) and the normal. The optimal M-estimator for the center of a t distribution with ν degrees of freedom is defined by the weight function $w(z) = \left(1 + z_i^2/\nu\right)^{-1}$ where $z_i = (y_i - \mu)/\sigma$; the weights depend on μ and σ through z_i. With this estimator, the influence of extreme observations is bounded, and falls to 0 far from μ. Note that the variance of the corresponding scaled t distribution is $(\nu/(\nu - 2))\sigma^2$, and is undefined for $\nu \leq 2$.

Another popular M-estimator is defined by the "Huber" ψ-function; see BSZ for results with this alternative.

The downside of robust estimation is that it may be biased if the population distribution is asymmetric. This issue is commonly avoided in robust distribution theory by postulating a symmetric error distribution, but this solution is not available to us in the survey context. There is no robust alternative to the usual unbiased estimators that guarantees unbiased estimation with all populations.

This presents us with a bias versus variance tradeoff. If the outlying observations are symmetrically distributed, the robust estimators may do well. If the outlying observations are asymmetrically distributed, however, the observations on the long-tailed side will be downweighted more on the average than those on the short-tailed side, which may make the estimator of location biased. By continuously varying the tuning parameter ν, we can investigate a range of alternatives from a sample mean with no downweighting (corresponding to $\nu = \infty$) to strong downweighting.

3.3 Downweighting Clusters Based on Their Infinitesimal Influence

Our downweighting strategy applies an M-estimator of the mean to influence statistics. The following steps are required: (i) we calculate the infinitesimal influence statistic D_i as given in (1) for each cluster in the sample; (ii) we calculate M-estimates of location and scale for the influence statistics; from this step, we (iii) retain the M-estimation weight for each cluster. Then (iv) we multiply the original weight for each cluster by the robust estimation weight, and (v) we recalculate the parameter (adjustment factor) for the poststratum using the new weights. Step (ii) is iterative; we also iterate the entire sequence of steps because the influence statistics calculated in (i) may themselves depend on the current estimates of the parameter.

We estimate the variance of the M-estimator using standard formulae (Huber 1981, eqn. 2-15). These formulae take into account the fact that the robust estimation weights are not known in advance but are estimated using estimates of μ and σ.

4 Robust Estimation with the 1990 Post-Enumeration Survey

In this section we apply the techniques described in Section 3 to the data on the clusters in the 1990 Post-Enumeration Survey. We use our M-estimator and explore the effect of various choices of the tuning constant ν. We compare the robust procedure to schemes that truncate large weights, and finally explore the issue of asymmetry of the distribution of influence statistics.

4.1 Net Undercount and the Influence Statistic

The adjustment factor in a poststratum is the ratio of estimated true population to the census count excluding substitutions,

$$A = (C/E)/(M/P), \tag{2}$$

where

$E = \sum W_{Ei} E_i$ = weighted estimate of total enumerations from E-sample,

$C = \sum W_{Ei} C_i$ = weighted estimate of correct enumerations from E-sample,

$M = \sum W_{Pi} M_i$ = weighted number of matches between the P sample and the census,

$P = \sum W_{Pi} P_i$ = weighted estimate of the population total from the P sample,

W_{Ei}, W_{Pi} are E- and P-sample weights respectively for cluster i, and E_i, C_i, M_i, and P_i are the corresponding unweighted counts of persons in cluster i. The above expression may be interpreted as the estimated fraction of census enumerations that are correct, divided by the estimated fraction of all persons in the poststratum who were enumerated in the census.

By taking derivatives of (2) with respect to the inclusion indicators for the clusters, we obtain the (infinitesimal) influence of cluster i on the estimator:

$$\begin{aligned} D_i \approx & -1/M[W_{Ei}(C_i - (C/E)E_i) - \\ & W_{Pi}(M_i - (M/P)P_i)], \end{aligned} \tag{3}$$

the approximation holding if $A \approx 1$ and the total number of PES matches in the poststratum is close to the number of correct enumerations. Note that a similar procedure may be applied for any statistic that is a function of estimated totals.

The two bracketed terms in (3) may be interpreted as the weighted excess of correct enumerations in the cluster over the expectation given E-sample size and the average correct enumeration rate, and the weighted excess of matches in the cluster over the expected number of matches given P-sample size and the average match rate. The influence therefore is approximately proportional to the excess of the weighted net number of cases contributed by the cluster to the estimated total undercount over the expectation for a cluster of that size. Thus, heuristically variance estimation using influence statistics is simply variance estimation for estimated net total undercounted persons in the poststratum: the influence-based variance estimate is roughly equivalent to what we would obtain if the point estimate were the sample total of net undercount by cluster.

4.2 Estimates from the 1990 Data

For an investigation of the effect of using robust estimators, we poststratified the PES clusters and calculated adjustment factors by poststratum. The poststratification variables are racial composition, tenure (owner versus renter) composition, and urbanicity (defined differently than in the standard PES poststratification); see BSZ for details. These variables define a 3×2×3 poststratification with 18 poststrata. Because our poststrata are defined by cluster characteristics, unlike those used in the actual 1990 PES estimation procedure, each cluster falls into only one poststratum. Hence to avoid making the poststrata excessively small, we did not use a fourth potential stratifier, the census division, in this analysis.

Our poststratification is by cluster rather than by person or household. This was imposed on us by the use of a cluster file rather than a microdata file for the analysis. It oversimplifies the analysis, because each cluster contributes to only one poststratum, so the poststratum estimates are essentially independent. (Extensions to a more realistic situation, with poststrata defined in terms of individual as well as cluster characteristics, are considered in Section 5.4).

To assess the distributional form of the influence statistics within poststrata, we first drew quantile plots (shown in BSZ). In every poststratum, the distribution was long-tailed relative to the normal distribution. A single quantile plot for all poststrata was created by z-scoring the influence statistics within each poststratum and then combining all the z-scores into a single distribution. The deviation in the tails of the normal plot from a straight line indicates a heavy-tailed distribution. We created t quantile plots for various values of ν. By eye, the best fit appeared to occur for $2 < \nu < 4$, and closer to 2 than to 4.

The maximum likelihood estimate of the degrees of freedom ν of a t distribution fitted to the combined z-scored influence statistics was 0.86. It would be quite disturbing if this were the underlying distribution of the influence statistics, since it would imply a distribution that is longer tailed than the Cauchy ($\nu = 1$) and has neither a variance nor a mean. The estimate $\hat{\nu}$ is very sensitive to a few particularly extreme observations, however, and we instead focus on a value $\nu = 2.5$ derived from the graphical investigation.

We calculated the robust estimate of the adjustment factor for each poststratum and the combined national data for the t-based ψ-function with $\nu = 100, 20, 8, 4$, and 2.5. To demonstrate the effects of varying the tuning parameter of the estimators, "trace plots" show the estimated undercount for each poststratum against the tuning parameter ν (Figure 1), with the unbiased survey-weighted mean at the left of each plot. (We use the term "unbiased" loosely to distinguish these estimates from the robust estimates, although because they are ratio estimates they are not strictly unbiased.) The heavy line toward the bottom represents the national undercount rate estimate. The poststrata whose estimates are most affected by downweighting appear as sharply rising or falling lines. The estimates for some poststrata change rapidly from the unbiased estimate ("Mean") to the $\nu = 20$ robust estimate, but move little beyond that point.

A similar plot for estimated standard errors (BSZ) shows that the estimated standard errors fall for every poststratum, and dramatically in a few, with a few exceptions at the smallest values of the tuning constant. The trace plots provide a graphical tool for considering the possible tradeoffs of bias against variance as the parameter which controls downweighting is varied.

Estimated undercount rates and their standard errors for the unbiased estimator and the "t" estimator with $\nu = 2.5$ are compared in Table 1. Of the 18 poststrata, the robust estimates for 7 differ from the unbiased estimates by over 1%. The largest difference is 4.0% for the Black suburban renter poststratum. Estimated standard errors for the robust estimator are from 15% to 88% as large those for the standard estimator; the average ratio is 39%.

4.3 Asymmetry and Bias

The discussion in Section 4.2 focused on considerations of variance and relative efficiency of estimators. As noted in Section 3.2, however, if the distribution of influence statistics is not symmetric, the robust estimators may be biased. If the bias were equal in every poststratum, it would be of relatively minor concern, because estimates of relative undercount would be unaffected. On the other hand, substantial differential biases in estimates would defeat the purpose of the entire undercount estimation program. We now consider the available evidence about bias.

Because of the balance designed into the PES, large geocoding errors should equally generate outlier undercount and overcount clusters. On the other hand, some other types of errors may not balance in this way, and therefore may generate a longer tail on one or the other side of the distribution of influence statistics. For example, there is no strong reason a priori to assume that high-weight clusters will contribute equally to extreme overcounts and undercounts.

In order to explore possible asymmetry of the influence statistic distributions, we prepared a quantile-quantile plot of the left side of the combined z-score distribution against the right side (split at the median). The left tail, corresponding to observations with positive influence on undercount estimates, is shorter than the right tail at the extremes, but there is an intermediate range at which the tail on the undercount side is heavier. This implies that the extreme observations will tend to be downweighted more on one side or the other (depending on the tuning parameter of the estimator), creating a bias.

One approach to determining whether downweighting of extreme clusters biases estimates of undercount rates is to ask whether the extremely influential clusters systematically affect differential undercount rates, or whether on the contrary these influential clusters are essentially randomly scattered among all poststrata. Differential bias is important because shares of population rather than absolute counts are critical to many uses of census data, such as apportioning representation or dividing up monetary benefits. A constant bias in undercount estimates might therefore have little importance.

We construct a randomization test whose null hypothesis is that the distributional form of the z-scores of influence statistics is the same in every poststratum. If this is true, the effect of applying downweighting is the same in every poststratum, on the scale of the z-scores. The alternative hypothesis is that the pattern of asymmetry differs between poststrata, so that the shifts due to downweighting are systematically different. The test statistic is the sum of the squared centered shifts, i.e., the squared differences between the mean z-score in each poststratum and the corresponding weighted mean after applying robust downweighting and centering the differences so their mean is zero across poststrata. The randomization distribution to which the observed value is referred is that obtained by z-scoring influence within each poststratum and then repeatedly randomizing the z-scores among the poststrata (without replacement) so that the number of clusters in each poststratum is the same as the number in the corresponding poststratum in the observed data set; 10,000 draws were taken for each test.

For the t estimator with $\nu = 2.5$, the observed value of the test statistic fell at the 93rd percentile of the randomization distribution. With $\nu = 20$, the test statistic fell at the 96th percentile, and with $\nu = 1000$ at the 95th percentile. Thus, although the the statistical evidence that the bias differs by poststratum is stronger for $\nu = 20$ or $\nu = 1000$ than for $\nu = 2.5$, the absolute magnitude of the changes in estimates in the former cases is smaller.

Another approach to exploring possible biases is to compare an estimate of bias, the difference between the unbiased and robust estimates, to the standard error of that difference. We estimate the standard error using the same Taylor linearization approach used for the adjustment factor, applied to the difference of two ratios. (We use an approximation that treats the robustifying weights as fixed, which probably slightly underestimates the variance.) Nationally (ignoring poststratification), downweighting with $\nu = 2.5$ reduces the undercount rate from 1.76 percent to 1.75 percent, an insignificant difference. We calculated the difference between poststratum and national undercount rate (relative undercount) for each poststratum. Although four relative undercounts change by more than 3 percent with $\nu = 2.5$ weighting, the t-statistics for testing significance of these changes were 2.09, 1.81, 1.75, and .97, and the largest of these was in a poststratum with only 11 clusters. Hence there is little evidence that these differences represent bias in the robust estimator rather than excess sampling error of the unbiased estimator.

Yet another approach compares the accuracy (MSE) of the unbiased and robust estimators in a way that allows us to aggregate across poststrata. We use the relationship $\mathrm{MSE}[y] - \mathrm{MSE}[x] = \mathrm{Var}[y] - \mathrm{Var}[y-x] + \mathbf{E}[(y-x)^2] - \mathrm{Var}[x]$, where x is an unbiased estimator and y is a possibly biased estimator (in this case, the robust estimator); the third term is estimated by the observed $(y-x)^2$. The estimated difference in MSE is negative for some poststrata and positive for others, but the unweighted average ($-2.45\%^2$) and population-weighted average ($-0.82\%^2$) of the independent estimates of the differences by poststratum are both negative, evidence that the MSE of the robust estimator is smaller. The largest negative term is from the "other renter rural" poststratum, where there is a large reduction in SE with only a modest change in the estimate; however, even excluding this poststratum the means are still negative ($-0.45\%^2$ and $-0.48\%^2$ respectively). The largest positive term is for the "Black renter suburban" poststratum; although the change in the undercount estimate is large, the robust estimate (4.97%) may be more plausible than the unbiased estimate

(0.95%), given the generally high undercount rates for renters and Blacks.

For a more powerful test of whether downweighting tends to differentially bias estimates for poststrata with high undercount rates, we tested the relationship between the change due to downweighting (with $\nu = 2.5$) and the estimated undercount rate. We fit a weighted linear regression, with weights inversely proportional to the estimated variance of the difference. Using the raw undercount rate as the predictor in the regression, the relationship appeared significant ($t = 3.99$); this effect would be expected, however, due to the correlation between the predictor and the error in the outcome variable. Using robust estimates of undercount rates as the predictor, there was weaker evidence ($t = 1.89$) of a relationship.

These analyses suggest that substantial decreases in variance can be achieved by downweighting, perhaps without a significant increase in the biases of poststratum estimates. Given the potential for bias due to asymmetry, it may be preferable to use an intermediate level of downweighting, like our example with $\nu = 20$, rather than a more drastic downweighting (e.g., using $\nu = 2.5$) that would be more nearly optimal for a symmetrical distribution.

4.4 Truncation of Weights

An alternative approach to robust estimation is to truncate weights of extreme clusters. This approach bears discussion for several reasons. First, the analysis of estimates using truncated weights offers insight into how much of the effect of the robust downweighting procedure described above can obtained by controlling weights alone. Second, truncation of weights might be less controversial than the influence-based procedure, if the downweighting is based purely on the design and not on the observed outcomes in each cluster. Finally, by investigating the variance of the truncated-weight estimators, we can evaluate the benefits that could be obtained by changing the design to avoid extreme weights.

We compared (BSZ) estimates and estimated standard errors nationally and by poststratum with the unbiased estimator, when both P- and E-sample weights are truncated at 2000 (affecting the weights in about 200 clusters for each sample), and when weights are made equal for all clusters. We find that in several poststrata (notably Other Rural Renters and Black Suburban Renters), the effect of truncating weights at 2000 is similar to that of robust estimation with $\nu = 20$: both methods lead to similar reductions in standard error and similar shifts in point estimates. In these poststrata, a redesign that avoided such extreme weights (which were largely due to the lower sampling rate for small blocks) would probably have similar effects on variance. In other poststrata (notably Hispanic Rural Renters and Hispanic Suburban Renters), truncation of weights at 2000 has very little effect even though the robust estimator substantially shifts the estimates and lowers the estimated standard error. This underlines the fact that extreme influence can result from either extreme weights or extreme unweighted net undercounts (or a combination of the two). Truncation of weights may be a more conservative procedure, but it also has less potential payoff.

5 Discussion

To conclude, we suggest some topics for future research related to the work in this paper in the context of census coverage estimation; these ideas are elaborated in BSZ.

5.1 Improved Sample Design and Processing

As suggested in Section 4.4, extreme weights contribute substantially to variance, especially in some poststrata. Some of these extreme weights appeared because of undersampling of small blocks, which in some cases turned out unexpectedly to have substantial populations and sometimes substantial undercounts or overcounts. It may be worthwhile in future coverage measurement efforts to select a larger sample of small blocks and screen them so that the ones that turn out to be heavily populated will not have such large weights.

Large-scale errors resulting from geocoding errors may be avoided by changing the procedures used in coverage measurement. If the search area had been extended for clusters with large errors, then some of the large nonmatched structures found under the 1990 design would have been matched in the extended area. On the other hand, our investigations suggested that other large-scale errors are not of a form that readily can be defined away by improvements in PES processing.

5.2 Non-Normality and Smoothing Models

Observations that are extreme and therefore influential on the mean also have extreme influence on variance estimates. If the data (or, more precisely, the cluster influence statistics) are normally distributed, then the sample mean and variance estimates are stochastically independent of each other. This is not the case when the data are t-distributed. Some of the well-known robustness of inference based on the t distribution stems from the fact that when there are extreme observations in the sample that greatly affect the mean, the variance estimate will also be inflated. If, however, variances are smoothed toward some model-based estimate, then this robustness is lost, even though the smoothed variance estimates may well be better than the unsmoothed estimates. This led to problems in empirical Bayes inferences under the normality assumption. We conjecture that better results might be obtained by using hierarchical models with t error structure (Liu and Rubin 1998), extending our robust approach to a hierarchical structure.

5.3 More on Asymmetry

The t-based downweighting algorithms described above give unbiased estimates with reduced variance under the assumption that the distribution of cluster influence statistics is symmetric in each poststratum. If this is not the case, downweighting may reduce variance but introduce an unknown bias.

In our example, it is not difficult to conceive of reasons why the distribution of influence statistics would not be symmetric and why this asymmetry might be systematically related to other characteristics of poststratification.

For these reasons, we suggest caution in the application of our methods until further research gives us a better way to characterize their effects on estimates. Several future extensions may extend the utility of these methods. For example, we might directly model the asymmetry of the influence distributions and thereby reduce the biasing effects of downweighting, by fitting truncated t distributions with possibly different scales and degrees of freedom to the two sides of the distribution. It may also be possible to estimate the downweighting parameters that give an optimal bias-variance tradeoff according to a criterion of estimated MSE as in Section 4.3.

5.4 Multivariate data

The estimation scheme of the PES is poststratified by characteristics of persons rather than of clusters (blocks). Consequently, more than one poststratum appears in each cluster, and both the observation and the influence statistic for each cluster are multivariate.

Several alternative extensions are possible to robustify estimation for multivariate estimands. If influence statistics for a cluster are independent for the different poststrata, it would make sense to calculate robust weights separately for each of the poststrata. A second strategy would be to calculate a single robust estimation weight, replacing the t model used above with a multivariate t distribution (Liu 1996). This would be sensible, for example, if we were estimating relationships among the variables and wanted to downweight values that are outliers from the usual relationships.

Given what we know about the reasons that some blocks are highly influential, the influence statistics of a block for estimates for several poststrata are likely to be dependent. Large errors that cause many households to be omitted or erroneously enumerated affect estimates in the same direction for several poststrata. Large weights similarly inflate the influence of households containing members from a number of poststrata, who are likely to be omitted or erroneously enumerated as a group. Consequently the influence of these blocks will usually have the same sign for a number of poststrata. Hence, they can be more sensitively and specifically detected by a measure of influence that sums across all poststrata, such as the influence of the block on the estimate of total population, or a weighted sum of influence on the population estimates by poststratum. Further research using data that break down undercount by poststratum within blocks will be required to develop these ideas.

5.5 Conclusion

We have explored the use of a robust estimator in a survey with influential clusters due to extreme observations and large weights. Despite the many unknowns, we believe that the large reductions in standard error suggested by Table 1 makes this is a promising area for future research.

The analyses presented here may help with the design of future coverage surveys to avoid the features that caused some clusters to be overly influential in the 1990 PES. Furthermore, even if there are some uncertainties about the properties of our estimators, a fairly good method that is prespecified and applied in an objective manner may be more useful and acceptable than one which is tailored to the data after it is collected.

References

Belin, T., G. Diffendal, S. Mack, D. Rubin, J. Schafer, and A. Zaslavsky (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussion). *Journal of the American Statistical Association 88*, 1149–1166.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association 81*, 1063–1069.

Church, J. and B. Harris (1970). The estimation of reliability from strength-stress relationships. *Technometrics 12*, 49–54.

Diffendal, G., A. Zaslavsky, T. Belin, and N. Schenker (1994). Influential observations in the 1990 post-enumeration survey. In *Proceedings of the 1994 Annual Research Conference*, Washington, D.C., pp. 523–548. U.S. Department of Commerce, Bureau of the Census.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions.* New York: John Wiley.

Hogan, H. (1993). The 1990 post-enumeration survey: Operations and results. *Journal of the American Statistical Association 88*, 1047–1060.

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*, 73–101.

Huber, P. J. (1981). *Robust Statistics.* New York: John Wiley.

Jaeckel, L. (1972). The infinitesimal jackknife. Technical report, Bell Laboratories, Murray Hill, New Jersey.

Lee, H. (1995). Outliers in business surveys. In B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge, and P. Kott (Eds.), *Business Survey Methods*, pp. 503–526. New York: John Wiley.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association 91*, 1219–1227.

Liu, C. and D. B. Rubin (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika 85*, 673–688.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the ASA Section on Survey Research Methods*, pp. 225–230.

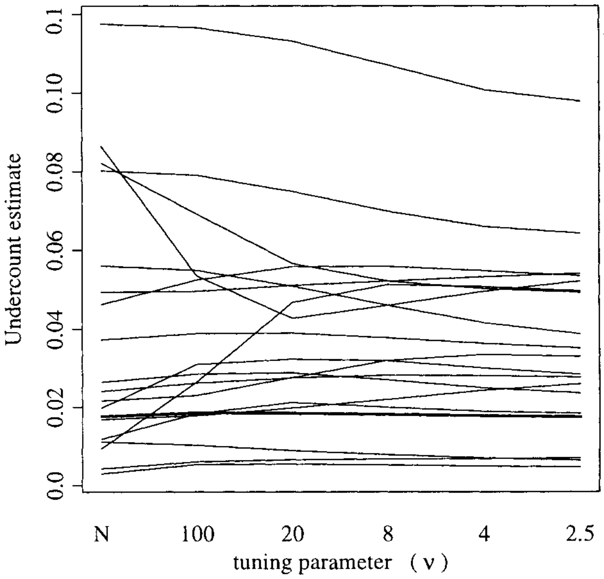

Figure 1: Traceplot of undercount estimates by poststratum and nationally (heavy line) against tuning parameter ν of robust estimator.

Table 1: Undercount estimates and estimated standard errors, by poststratum and nationally, for no downweighting (normal) and downweighting based on the t distribution with 2.5 degrees of freedom ($t_{2.5}$). "SE%" gives the robust standard error as a percentage of that of the conventional estimator.

Poststratum	Normal		$t_{2.5}$		
	UC	SE	UC	SE	SE%
Black Rural Owner	5.61	1.47	3.87	0.87	59%
Black Suburban Owner	1.19	1.34	1.85	0.45	34%
Black Urban Owner	3.72	0.70	3.51	0.31	44%
Black Rural Renter	11.75	2.62	9.78	1.17	45%
Black Suburban Renter	0.95	2.08	4.97	0.58	28%
Black Urban Renter	4.63	1.48	5.36	0.46	31%
Hispanic Rural Owner	2.16	1.27	3.30	0.77	61%
Hispanic Suburban Owner	2.64	0.85	2.37	0.37	44%
Hispanic Urban Owner	2.41	0.59	2.78	0.28	47%
Hispanic Rural Renter	8.02	2.33	6.43	2.06	88%
Hispanic Suburban Renter	8.22	2.19	4.93	0.55	25%
Hispanic Urban Renter	4.94	0.89	5.41	0.38	43%
Other Rural Owner	0.43	0.41	0.72	0.15	37%
Other Suburban Owner	1.12	0.32	0.65	0.09	28%
Other Urban Owner	0.30	0.47	0.49	0.10	21%
Other Rural Renter	8.64	5.00	5.22	0.77	15%
Other Suburban Renter	1.98	1.64	2.84	0.27	16%
Other Urban Renter	1.68	0.75	2.60	0.35	47%
National	1.76	0.22	1.75	0.06	27%