

Downweighting Influential Clusters in Surveys, with Application to the 1990 Post-Enumeration Survey

Alan M. Zaslavsky, Nathaniel Schenker, and Thomas R. Belin *

December 17, 1999

*Alan M. Zaslavsky is Associate Professor, Department of Health Care Policy, Harvard University, Boston, MA 02115 (email: zaslavsk@hcp.med.harvard.edu); Nathaniel Schenker is Senior Scientist for Research and Methodology, National Center for Health Statistics, Hyattsville, MD 20782 (email: nhs1@cdc.gov); and Thomas R. Belin is Assistant Professor in Residence, Departments of Psychiatry and Biostatistics, UCLA, Los Angeles, CA 90024 (email: tbelin@mednet.ucla.edu). 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.

Abstract

Certain primary sampling units (PSUs) may be extremely influential on survey estimates and consequently contribute disproportionately to their variance. We propose a general approach to estimation that downweights highly influential PSUs. Our robust estimation strategy applies M-estimation to the empirical influence of the PSUs. The method is motivated by a problem in census coverage estimation, and we illustrate it using data from the 1990 Post Enumeration Survey. In this context, an objective, prespecified methodology for handling influential observations is essential to avoid having to justify judgemental post hoc adjustment of weights. In 1990, both extreme weights and large errors led to extreme influence. We analysed these data using M-estimators based on the t distribution and the Huber ψ -function, and estimated influence by Taylor linearization of the estimator. As predicted by theory, the robust procedures greatly reduced 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 consider the properties of the estimators in the presence of asymmetry and demonstrate techniques for assessing the bias-variance tradeoff, finding that mean squared error is reduced by applying the robust procedure to our dataset. We also suggest PES design improvements to reduce the impact of influential clusters.

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

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 and correctly recorded (“nonrepresentative” and “representative” outliers). Here we presume that the data are accurate.

Previous research on controlling influential observations in surveys deals with one of two problems: outlying data values or extreme weights. All methods for handling these problems entail trading possibly increased bias for diminished variance to reduce mean squared error (MSE).

One strategy for handling outliers is to identify and edit them. Chernick and Murthy (1983) used the influence function (Hampel 1974) to detect outliers and edit them by replacing them with values having little or no influence on target estimates. Smith (1997) and Smith, Kocic, and Hibbitt (1997) identified outliers in an economic survey using an influence statistic and Winsorized outlying observations to agree more closely with those predicted by a model. Related contributions include Chernick, Downing, and Pike (1982) for time-series data and Little and Smith (1987) for multivariate data.

A second strategy for dealing with outliers is to apply robust estimation techniques to the data. For example, Chambers (1986) and Gwet and Rivest (1992) bounded the influence of extreme values by applying the M-estimation approach of Huber (1964). Hulliger (1995) applied M-estimation to develop a robust version of the traditional Horvitz-Thompson estimator. Hidiroglou and Srinath (1981) and Chambers (1986), among others, have shown that substantial reductions in MSE are possible by downweighting outliers.

A common strategy for handling extreme weights in surveys is weight trimming or truncation (Potter 1988; Potter 1990), that is, imposing a ceiling on weights. Potter (1990) explored various alternatives for identifying a ceiling and offered recommendations. Stokes (1990) compared weight truncation with a theoretically motivated strategy for shrinking weights in a stratified sampling context and found little basis for preferring one method to the other.

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 sensitivity of the estimator to inclusion of the cluster in the data. A new estimator is then calculated using modified weights, where the modification factors are determined by fitting a long-tailed (e.g. t) distribution to the influence statistics.

Our approach differs from the robust estimation approaches cited above as well as other applications of t error models in the statistics literature (e.g., West 1984; Lange, Little, and Taylor 1989; Lange and Sinsheimer 1993; Liu 1996; Fernandez and Steel 1999), 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 an outlying data value but a low weight

might be moderately influential. A robust method applied directly to the raw data values would downweight such a cluster severely, and weight trimming might not downweight the cluster at all. In contrast, our method would downweight the cluster moderately.

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 robust procedure substantially reduces 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 in 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 persons 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 robust 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 (independent of the census) 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.

The sample design gave special consideration 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. 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.

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). E-sample search rules were similar. Consequently, 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 E-sample 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 outside the search area and therefore classified as erroneously geocoded.

A small percentage of individuals in the P sample and E sample did not have resolved enumeration statuses after PES processing. For these individuals, probabilities of being included in the census (for P-sample cases) or of being correct enumerations (for E-sample cases) were imputed from models fitted to data from those whose enumeration status was resolved (Belin, Diffendal, Mack, Rubin, Schafer, and Zaslavsky 1993). These probabilities were then used in the dual-system estimator to be described below.

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. Population estimates for poststrata were based on the dual-system estimator of the population,

$$\text{DSE} = (\text{CEN} - \text{SUB}) \frac{C/E}{M/P}, \tag{1}$$

where CEN is the raw census count, SUB is the number of people imputed by substituting complete households for households for which data were not obtained, C is the weighted estimate of correct enumerations from the E sample, E is the weighted estimate of total enumerations from the E sample, M is the weighted number of matches between the P sample and the census, and P is the weighted estimate of the population total from the P sample. The first factor (CEN–SUB) represents the actual census enumerations, since substitutions obviously could not be matched against the PES. The numerator C/E adjusts total census enumerations to correct enumerations, and the denominator M/P adjusts counts of enumerated persons to estimated total population.

The “adjustment factor,” $(C/E)/(M/P)$, was calculated for each poststratum. It is the key estimand because it is used for the adjustment of small area populations. In the 1990 PES, there were 1,392 poststrata, and the corresponding 1,392 adjustment factors were subsequently smoothed using a hierarchical regression model. Further details on 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, see Diffendal, Zaslavsky, Belin, and Schenker (1994).

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, that is, weights that are very large compared to weights for other clusters in similar areas or with similar population groups. 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.

As an example of an error with high influence, a geocoding discrepancy in one block in the Southeast United States led, under PES matching rules, to an estimated undercount of approximately 75% in that block, which weighted up to approximately three-quarters of a million people. An apparent overcount of similar magnitude occurred due to a geocoding error in a block in the Northeast. These large and opposite errors made a large contribution to estimated differential undercount between the Southeast and Northeast that was apparently determined by the accident that the P sample in the Southeast contained a cluster where a large number of misgeocoded people actually lived whereas the E sample in the Northeast contained a cluster where such people were mistakenly placed. The errors could have just as easily gone the other way in both regions, with an equal impact on differential undercount but in the opposite direction. In fact, a somewhat arbitrary decision was made to reduce the weight of these blocks so that the impact of each on the undercount estimate was 150,000 persons. In reprocessing for a later set of estimates, one of the errors was eliminated by extending the search area.

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. Therefore, downweighting

influential clusters, as was done in 1990, is appropriate. Post-hoc downweighting is subject to criticism, however, because it relies on expert judgement applied to individual blocks and because of the discontinuity between those blocks selected for downweighting and those not selected. Our objective in Sections 3 and 4 is to provide an objective and systematic basis for downweighting.

2.4. Patterns of Influential Clusters

Diffendal, Zaslavsky, Belin, and Schenker (1994) gave a detailed report on an exploratory analysis of a cluster-level file containing gross undercount, gross overcount, net undercount, average P-sample and E-sample weights, the extent of certain specific types of errors, demographic characteristics of cluster residents, and some other cluster-level characteristics for each of the 5,290 clusters in the 1990 PES. Certain PES errors, such as a computer coding error that misclassified some correct enumerations as erroneous (Mulry and Spencer 1993), had been corrected in the cluster-level file. Moreover, the file reflected the determinations from an additional review of the 100 clusters that contributed most to undercount or overcount, which was undertaken as part of the 1992 postcensal estimates review program. The search area was extended in this operation for some cases to reduce the effects of large-scale geocoding errors. Thus there were fewer influential clusters in this analytic file than in data used for 1990 undercount estimation.

Diffendal, Zaslavsky, Belin, and Schenker (1994) examined relationships between cluster characteristics and the influence of clusters on the estimate of national net undercount. (Although Section 4 considers influence on poststratum undercount rates, the distinguishing patterns would be similar in either analysis.) Briefly, they found the following relationships. Small blocks appeared to contribute disproportionately to clusters being influential in terms of large undercount. Small blocks were more likely to make a moderate contribution to overcount than a moderate contribution to undercount. Large blocks that were subsampled also had big weights that often made them influential. Generally, higher weights were associated with greater influence, as were large discrepancies between the average E-sample and P-sample weights within the same cluster.

Several categories associated with high levels of census error in general also had disproportionate numbers of influential clusters. With regard to geography, the highest percentages of influential clusters (high undercounts or overcounts) occurred in the Middle Atlantic, South Atlantic, and Pacific Divisions. The South Atlantic and Pacific Divisions had unusually high percentages

of clusters with high undercounts. Further investigation is needed to see whether the lack of homogeneity in influence was due to cluster-level errors or simply varying general levels of coverage error in these divisions. There were more influential clusters in urban and rural areas than there were in suburban areas.

Perhaps surprisingly, although higher undercounts were associated with higher percentages of minority residents, the overall percentage of influential clusters (high undercount or overcount) appeared to be highest in clusters with medium percentages (10% to 50%) of minority residents. By tenure, the highest percentage of influential clusters occurred for the category with the highest proportion of renters (and the fewest owners). This could be explained by the fact that renters experienced high error rates in general and are more likely to live in large buildings that might be subject to large-scale errors.

3. ROBUST ESTIMATION WHEN THERE ARE INFLUENTIAL CLUSTERS

We now present models and theoretical perspectives that suggest methods to reduce the contribution of influential clusters to variance and thereby potentially to improve the quality of survey estimates. In short, the variance of an estimator is often approximately equal to the variance of the sample mean of the influence of the observations. This suggests that the variance of the estimator might be reduced by reweighting the observations, downweighting those with extreme influence, using weights derived by applying techniques for robust estimation of the mean of the influence values.

In Section 3.1, we discuss the influence function and its relationship to the variance of an estimator, and we define an influence measure for complex sample surveys. In Section 3.2, we discuss robust estimation using M-estimators. Finally, in Section 3.3, we discuss how we use our influence measure and M-estimation to downweight influential clusters, and how we estimate the variance of the resulting survey estimate.

3.1. Influence and Variance

Let x_1, \dots, x_n be independent and identically distributed observations of a random quantity X , where X has distribution function F , and let F_n denote the empirical distribution function, which assigns mass $1/n$ to each x_i . Consider an estimator $\hat{\theta} = t(F_n)$ of $\theta = t(F)$. The influence function

(Hampel 1974) is defined by

$$U(x, F) = \lim_{\epsilon \rightarrow 0} \frac{t[(1 - \epsilon)F + \epsilon\delta_x] - t(F)}{\epsilon} \quad (2)$$

where δ_x is a point mass at x . The influence function can be regarded as measuring the sensitivity of the estimator to inclusion of the data value x . Under suitable regularity conditions (Hampel 1974),

$$t(F_n) = t(F) + n^{-1} \sum_{i=1}^n U(x_i, F) + O_p(n^{-1}), \quad (3)$$

It follows from expansion (3) that

$$\text{Var}(t(F_n)) \doteq \text{Var}(n^{-1} \sum_{i=1}^n U(x_i, F)), \quad (4)$$

that is, the variance of the estimator is approximately equal to the variance of the sample mean of the influence of the observations.

Various approximations to the influence of an observation, $U(x_i, F)$, lead to various related estimates of variance based on approximation (4); see Efron (1982) for a detailed discussion. If we set $\epsilon = -1/(n - 1)$ and $F = F_n$ in (2) without taking the limit, we obtain the approximation $\hat{\theta}_i - \hat{\theta}$, where $\hat{\theta}_i$ is the jackknife pseudo-value (Huber 1981). This provides a motivation for jackknife variance estimation. The empirical influence component, $U(x_i, F_n)$, is obtained by letting $F = F_n$ in (2) and taking the limit. This is a motivation for the infinitesimal jackknife estimate of variance (Church and Harris 1970; Jaeckel 1972), $n^{-2} \sum_{i=1}^n U^2(x_i, F_n)$, which equals the Taylor linearization (delta method) estimate of variance if $\hat{\theta}$ can be expressed as a function of means (Efron 1982).

The concept of influence and the related approximations and variance estimates can be extended to the context of survey sampling. Consider first a single-stage (element or cluster) with-replacement sampling design. For unit i ($i = 1, \dots, N$) in the population, let x_i , w_i , and I_i denote, respectively, the variables, sampling weight (inverse of probability of selection), and inclusion indicator ($I_i = 1$ if unit i is included in the sample, $I_i = 0$ otherwise). Then the (x_i, w_i) for the n units in the sample may be viewed as n independent and identically distributed observations from a distribution G that assigns mass $1/w_i$ to unit i in the population. The associated extension of the empirical influence of a unit in the sample is thus $U(x_i, w_i, G_n)$, where G_n assigns mass $1/n$ to each unit in the sample. Note that G_n maps to F_n , the weighted empirical

distribution of the data for the sampled elements in the clusters, so we still define $\hat{\theta} = t(F_n)$ and can calculate the empirical influence as $U(x_i, w_i, F_n)$.

For a multistage sampling design, the further extension of $U(x_i, w_i, G_n)$ is immediate, with x_i now representing the data and lower-stage weights (or sufficient statistics, if they exist) for primary sampling unit i , and w_i now being the first-stage weight. Variance estimates based on $U(x_i, w_i, G_n)$, such as the infinitesimal jackknife estimate, apply to sampling without replacement if finite population corrections can be ignored. For analogous jackknife approximations and modifications for stratified sampling, see Wolter (1985, ch. 4) and Lohr (1999, ch. 9).

The empirical influence $U(x_i, w_i, F_n)$ is particularly simple to calculate when we estimate a function of the population mean $\theta = f(\bar{X})$ by applying the function to the weighted sample mean, $\hat{\theta} = f(\bar{x})$, where $\bar{x} = (\sum_{i=1}^N I_i w_i x_i) / (\sum_{i=1}^N I_i w_i)$. From (2), using the fact that the sample mean is a linear function of the empirical distribution, we have

$$U(x_i, w_i, F_n) = (w_i / \bar{w}) f'(\bar{x})(x_i - \bar{x}), \quad (5)$$

where $\bar{w} = \sum_{i=1}^N I_i w_i / n$ is the sample mean of weights and the product $f'(\bar{x})(x_i - \bar{x})$ is an inner product if x_i is multivariate. Thus, the empirical influence reflects the outlyingness of x_i (in a particular direction) and its relative weight. Similar arguments apply for other estimators that can be approximated as a weighted mean or sum, as when a maximum likelihood estimate is calculated using a Newton-Raphson step that involves a sum of score statistics.

The derivative-based influence statistic has appeal when it has a simple and possibly interpretable closed form. This is true in our application (Section 4.1), where the estimator is a function of means and therefore can be calculated using (5). A jackknife influence statistic could be used in other situations.

Note that the weights enter the calculation of influence in two ways. In general, they appear directly through the weight w_i of the observation, and indirectly through estimation of F_n , the distribution at which the derivatives are evaluated. In the case of a function of means, the latter appears through the value of \bar{x} in the factor $f'(\bar{x})$.

3.2. Long-Tailed Distributions and M-Estimation

It is well known that the sample mean is the optimal estimator of location for a normally distributed population but can be inefficient for symmetric distributions with heavier tails, due to its

sensitivity to outliers. M-estimators (Huber 1964; Huber 1981; Hampel, Ronchetti, Rousseeuw, and Stahel 1986) comprise a large class of robust estimators. For a location-scale family with location μ and scale σ , M-estimators are defined by solving estimating equations of the form $\sum_i \psi((y_i - \mu)/\sigma) = 0$ and $\sum_i \chi((y_i - \mu)/\sigma) = 0$ for μ and σ . If the estimating equations can be obtained by equating the derivatives of a log-likelihood to zero, then the M-estimator is a maximum likelihood estimator (MLE), but the usefulness of M-estimators does not depend on whether a specific distributional assumption holds or even whether the estimator is an MLE.

Typically, M-estimators are calculated iteratively. In particular, the M-estimator of μ may be calculated as an iteratively weighted mean, $\hat{\mu} = \sum_i d_i y_i / \sum d_i$, where $d_i = \psi(z_i)/z_i$ and $z_i = (y_i - \mu)/\sigma$ depends on the previous estimate of μ and σ . Robust M-estimators, such as those based on assuming a long-tailed distribution, give reduced weight to the extreme observations. This downweighting is the source of the robustness of the estimator against outliers.

The t family is useful for defining an M-estimator 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 $d_i = d(z_i) = 1 / (1 + z_i^2/\nu)^{-1}$. 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 $\nu \leq 2$.

Under the assumption that the sample is actually drawn from a t_ν distribution, the asymptotic relative efficiency (ratio of variances) of the optimal weighted estimator relative to the sample mean is

$$\text{ARE} = \frac{\nu(\nu + 1)}{(\nu - 2)(\nu + 3)}. \tag{6}$$

For example, with $\nu = 3$, ARE=2; with $\nu = 6$, ARE=1.17; and with $\nu = 15$, ARE=1.03. Thus, with heavy-tailed distributions, substantial gains in efficiency are possible by downweighting of outliers, although gains are modest with distributions that are fairly close to the normal distribution.

Another popular M-estimator can be defined by assuming a density that is normal in the middle and joined smoothly and symmetrically to exponential tails. This leads to the the ‘‘Huber’’

ψ -function,

$$\psi(z) = \begin{cases} -c, & z < -c \\ z, & -c < z < c \\ c, & z > c. \end{cases}$$

The corresponding weight function has $d(z) = 1$ for $|z| \leq c$ and $d(z) = c/|z|$ otherwise. As $c \rightarrow 0$, the location estimator approaches the median. An appealing property of this robust estimator is that observations in the tails all have the same influence on it.

A disadvantage 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 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 or c , we can investigate a range of alternatives from a sample mean with no downweighting (corresponding to $\nu = \infty$ or $c = \infty$) to strong downweighting.

3.3. Downweighting Influential Clusters

Our general approach is to downweight influential clusters based on weights derived by applying M-estimation of location to the influence statistics for the clusters. The expansion (3) suggests that a robust estimator of $\theta = t(F)$ can be obtained by applying M-estimation to influence statistics. Instead of relying solely on approximation (3), however, we use the robust modification weights d_i from M-estimation to recalculate F_n , together with sampling weights w_i , to recalculate F_n . Hence we iteratively recalculate both $\hat{\theta} = t(F_n)$ and the influence statistics. This leads to the following iteration:

Initially, set $d_i = 1$, $i = 1, \dots, N$. Define F_n^* as the empirical distribution based on weights $w_i d_i$. Then iterate the following steps until convergence:

- (i) Calculate the empirical influence statistic $u_i = U(x_i, w_i, F_n^*)$ for each cluster in the sample.

(ii) Calculate M-estimates of location and scale for the influence statistics $\{u_i\}$. Retain the final M-estimation weights, $\{d_i\}$.

(iii) Update F_n^* using the new values of $\{d_i\}$, and calculate $\hat{\theta} = t(F_n^*)$.

The reason we iterate steps (i) to (iii) is that the “data” $\{u_i\}$ used in M-estimation depend on the current values of the updated weights $\{d_i\}$ calculated in step (ii), through the recalculation of F_n^* . In particular, if θ is a function of means, then the influence statistics depend on a derivative evaluated at the mean, as noted at the end of Section 3.1. Step (ii) is typically iterative as well, so there are two layers of iteration.

The robust estimator $t(F_n^*)$ has the same form as the standard survey estimator $t(F_n)$ except that the weights have been modified. To estimate the variance of our final survey estimate after downweighting the influential clusters, it would be tempting to apply a standard survey-based procedure as if the modified weights were predetermined survey weights. This would be inappropriate, however, since the modification factors in the weights are estimated and depend on the empirical influence statistics, which themselves are estimates. A straightforward solution is to use a replication method such as the jackknife or the bootstrap, with the modification factors re-estimated for each replicate. We use the jackknife in this paper.

An alternative approach that yields a closed-form solution is to use empirical versions of standard formulae (Huber 1981, p. 45, equations 2.14 and 2.15) for the asymptotic variance of an M-estimator of location, which through (3) approximates the variance of $\hat{\theta}$. The resulting variance estimate is

$$\sum_{i=1}^n \psi(z_i)^2 / \left(\sum_{i=1}^n \psi'(z_i) \right)^2, \quad (7)$$

where z_i are the final z -scores of the influence statistics used in the M-estimation (step (ii)) of our iterative procedure. This approach will tend to underestimate variance somewhat. While it accounts for variability due to M-estimation of location, it does not account for variability due to M-estimation of scale and due to estimating influence by the empirical influence. Section 4.2 compares this approach with the jackknife approach.

4. ROBUST ESTIMATION WITH THE 1990 POST-ENUMERATION SURVEY

In this section we apply the techniques described in Section 3 to the data described in Section 2.4 on the clusters in the 1990 Post-Enumeration Survey. We use t -based and “Huber” M-estimators and explore the effect of various choices of the tuning constant (ν or c). We compare our robust estimation 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

Recall from Section 2.2 that the adjustment factor in a poststratum is the estimated ratio of true population to the census count excluding substitutions, estimated as

$$A = (C/E)/(M/P), \quad (8)$$

where

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

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

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

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

and E_i , C_i , M_i , and P_i are the unweighted counts of persons in cluster i . The E- and P-sample weights W_{Ei} , W_{Pi} incorporate first-stage weights w_i and further weights for subsampling within clusters, if any, which can differ for the E and P samples. Expression (8) represents 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.

Because (8) is a function of estimated totals (or equivalently weighted means) of $x_i = (C_i, E_i, M_i, P_i)$, we calculate by (5) the empirical influence of cluster i on the estimator:

$$\begin{aligned} u_i &= (A/\bar{w})(W_{Ei}C_i/\bar{C} - W_{Ei}E_i/\bar{E} - W_{Pi}M_i/\bar{M} + W_{Pi}P_i/\bar{P}) \\ &= nA[W_{Ei}(C_i - (C/E)E_i)/C - W_{Pi}(M_i - (M/P)P_i)/M] \\ &\approx (n/M)[W_{Ei}(C_i - (C/E)E_i) - W_{Pi}(M_i - (M/P)P_i)], \end{aligned} \quad (9)$$

where \bar{w} is the mean of first-stage weights and the estimated total is $C = n\bar{w}\bar{C}$ (and likewise for $E, M,$ and P). The final approximation holds if A is close to 1 and the total number of PES matches in the poststratum is close to the number of correct enumerations, as is generally the case.

The two terms square brackets in (9) can 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.

The current estimate of the empirical distribution appears in the algorithm of Section 3.3 through the ratios C/E and M/P in (9). Hence these ratios must be recalculated at each iteration in step (iii), using modification weights from robust estimation.

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). The cutoffs for these variables are shown in Table 1. These variables define a $3 \times 2 \times 3$ poststratification with 18 poststrata (Table 1). To avoid making the poststrata excessively small, we did not use a fourth potential stratifier, the census division, in this analysis.

Table 1 about here

Our poststrata are defined by cluster characteristics, rather than by person or household characteristics as in the 1990 PES estimation procedure, because only a cluster file rather than a microdata file was available for our analysis. Poststratification by cluster simplifies the analysis because each cluster falls into only one poststratum, so the poststratum estimates are essentially independent. Extensions to a more realistic situation, with poststrata defined by individual as well as cluster characteristics, are considered in Section 5.4. Furthermore, stratum indicators

were removed from the file, although the weights induced by the stratification were included, and hence we analyzed the data as if they were from an unstratified unequal-probability sample.

We express our empirical results in terms of the undercount rate $1-1/A = 1-1/(1+(A-1)) \approx A-1$, because this is the more familiar metric for describing undercount. For each poststratum, we calculated the estimated undercount with no downweighting and the estimated standard error (Table 2, columns labeled “Normal”). Standard errors were estimated using the jackknife.

Table 2 about here

To assess the distributional form of the influence statistics within poststrata, we first drew normal quantile plots. In every poststratum, the distribution was long-tailed relative to the normal distribution. A single normal 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 combined plot (Figure 1) is similar to the plots for individual poststrata. Normally distributed data would give an approximately straight plot; the deviation in the tails indicates a heavy-tailed distribution.

Figure 1 about here

To fit a t distribution to the influence statistics, 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. Figure 2 shows a combined t quantile plot of the z -scored influence statistics for $\nu = 2.5$. The plot is clearly much straighter than Figure 1, although the fit is still imperfect in the extreme tails.

Figure 2 about here

The maximum likelihood estimate of the degrees of freedom 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 the value $\nu = 2.5$ derived from the graphical investigation. A similar graphical investigation suggests

a value of c between .5 and 1 for the Huber ψ -function, although the exponential tails of the corresponding distribution are not heavy enough to match the tails of the influence distribution.

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 3), with the asymptotically unbiased survey-weighted estimate at the left of each plot. (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 to the $\nu = 20$ robust estimate, but move little beyond that point. Examination of the corresponding plot for estimates using the Huber ψ -function with $c = 8, 4, 2, 1, .5,$ and $.25$ (not shown) similarly shows that the “Huber” estimates change little for $c \leq 1$.

A similar plot for estimated standard errors (Figure 4) shows that the estimated standard errors fall for every poststratum but one, and dramatically in a few, with a few exceptions at the smallest values of the tuning constant. The exception is the Hispanic rural renter stratum, which had only 11 blocks so the sampling variance estimates are unreliable. The trace plots provide a graphical tool for considering the possible tradeoffs of bias against variance as the parameter which controls downweighting is varied.

Figure 3 about here

Figure 4 about here

Estimated undercount rates and their standard errors for the unbiased estimator, the t -based estimator with $\nu = 2.5$ and the “Huber” estimator with $c = .5$ are compared in Table 2. We first compare the unbiased and t -based robust estimates (Figure 5), and the corresponding standard errors (Figure 6). Of the 18 poststrata, the robust estimates for 8 differ from the unbiased estimates by over 1%. The largest differences are +3.6% for the Black suburban renter poststratum, and -3.6% for Hispanic suburban renters. Estimated standard errors for the robust estimator are from 23% to 87% as large as those for the standard estimator, except for the Hispanic rural

reuter poststratum for which the estimated standard error is 191% as large, based on 11 blocks; the average ratio (across all 18 poststrata) is 56% (45% if weighted by sample size).

Figure 5 about here

Figure 6 about here

The ARE from (6) of the robust estimator for $\nu = 2.5$ relative to the unbiased estimator, under the corresponding $t_{2.5}$ model, is 3.18. This implies that standard errors of the robust estimator would be about 56% as large as those for the approximately unbiased estimator, agreeing with the mean value in Table 2. Thus, substantial variance reductions appear to be possible.

The two robust estimators yield very similar point estimates and standard errors. Estimated standard errors for the t estimator range from 82% to 118% of those for the Huber estimator, averaging 103% as large. The differences between the robust estimates range from $-.4\%$ to $+.8\%$, and the mean absolute difference is only $.2\%$. In theory, the two robust estimators represent slightly different views of the influential observations. If we believe that blocks are highly influential due to a process (gross errors) which is distinct from and uninformative about the net undercount for the poststratum, we might want to ignore influential blocks altogether. The t -based estimator has the corresponding property of declining influence in the tails. On the other hand, if we believe that influential blocks are informative about undercount (as are “small blocks” with extreme weights) but should not be allowed too much influence on the estimates, we might prefer the Huber estimator, which bounds but does not eliminate the influence of these blocks. In practice, however, it appears that the number of influential blocks is small enough that the difference between the estimators is unimportant.

The asymptotic variance estimator (7) moderately underestimates the standard error, obtaining on the average 96% of the jackknife estimate of standard error for the survey-weighted estimator and 89% of the jackknife estimate for the robust ($\nu = 2.5$) estimator (from 60% to 128% for individual poststrata).

4.3. Asymmetry and Bias

Section 4.2 compared the variances and relative efficiencies of estimators. As noted in Section 3.2, however, if the distribution of influence statistics is not symmetric, the robust estimators may

be biased. Equal biases in every poststratum are of relatively minor concern in the undercount estimation program, 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. Large differential biases, on the other hand, would defeat the purpose of undercount estimation, which is to remove such biases. We now consider the evidence about bias.

Because of the balance designed into the PES, large geocoding errors should generate equal numbers of outlying clusters contributing to undercount and overcount. 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 (Figure 7) of the left side of the combined z -score distribution against the right side (split at the median). As in Figure 2, the influence is evaluated here at the initial estimate of the distributions and does not reflect the iterative recalculation of $\{u_i\}$ (Section 3.3). The left tail, corresponding to observations with negative influence on undercount estimates, is heavier than the right tail over part of the range but lighter at an intermediate range. 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), possibly creating a bias.

Figure 7 about here

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.

Section 2.4 presented a brief exploratory and descriptive analyses of the relationship between extreme influence and characteristics of the cluster that enter into poststratification. Here we address differential bias through a hypothesis test.

We construct a randomization test whose null hypothesis is that the distributional form of the 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 variance of the shifts, i.e., of the differences between the mean z -score in each poststratum and the corresponding weighted mean after applying robust downweighting. 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 16th percentile of the randomization distribution. Similarly with $\nu = 10, 20$, or 1000, the test statistic fell near the middle of the distribution. Hence this test gives little evidence of differential bias.

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 (estimated with the jackknife) of that difference. Nationally (ignoring poststratification), downweighting with $\nu = 2.5$ reduces the undercount rate from 1.76% to 1.65%, an insignificant difference. We calculated the difference between poststratum and national undercount rate (relative undercount, the most policy-relevant measure) for each poststratum, the amount by which each estimated difference changed with t -based robust estimation using $\nu = 2.5$, and the standard error of that change. The largest t -statistic for testing significance of these changes was 2.43 for a poststratum that was shifted by only .66%, and the others were less than 2; the sum of the squared t -statistics was 19.1, only slightly larger than its expectation. 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 $\text{MSE}[y] - \text{MSE}[x] = \text{Var}[y] - \text{Var}[y-x] + \text{E}[(y-x)^2] - \text{Var}[x]$, 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 ($-3.65\%^2$) and population-weighted average ($-1.10\%^2$) 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 Rural Renter” poststratum, where there is a large reduction in SE with only a modest change in the estimate; however, if we exclude this poststratum the means are still negative ($-1.10\%^2$ and $-0.64\%^2$ respectively). The largest positive term is for the “Black Suburban Renter” poststratum; although the change in the undercount estimate is large, the robust estimate (4.53%) is more plausible than the unbiased estimate (0.95%), in light of 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. Regardless of whether the unbiased or robust estimate of the undercount rate was the regressor, there was no significant evidence of a trend.

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

4.4. Truncation of Weights

An alternative approach to robust estimation is to truncate extreme sampling weights. 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 be 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.

Table 3 compares estimates and estimated standard errors (estimated using the jackknife) nationally and by poststratum with the unbiased estimator, when weights are truncated so that both P- and E-sample weights are no more than 2000 (affecting the weights in 803 clusters), and

when weights are made equal for all clusters. Weight truncation had effects similar to those of robust estimation on point estimates in several poststrata with large shifts, notably Black Suburban Renters and Other Rural Renters, but not in others. Likewise, in several poststrata truncation reduced standard errors substantially, but in some others standard errors with truncated weights are larger than those with the unbiased estimator. In every poststratum, the robust procedures afforded a greater reduction in variance than either weight truncation scheme. 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.

Table 3 about here

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 include ideas regarding sample design and processing, possible relations of our work to smoothing of undercount estimates, modifications of our procedures that could alleviate the problems caused by asymmetry, and extensions to multivariate estimands.

5.1. Improved Sample Design and Processing

The analyses in Section 4 demonstrate that extreme weights and large-scale errors can greatly increase variance. Because a relatively small number of clusters contribute disproportionately to these problems, it may be cost-effective to develop programs that attempt to minimize and ameliorate both sources of variance.

As suggested in Section 4.4, extreme weights contribute substantially to variance, especially in certain poststrata. Some of these extreme weights appeared because of undersampling of small blocks, which occasionally turned out unexpectedly to have substantial populations and sometimes had 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 found to be heavily populated will not have such large weights. For the 2000 Census, the Census Bureau changed its rules for initial listing to allow small blocks to be clustered with adjacent

larger blocks, reducing the number of small block clusters nationwide from nearly 3 million to just over 1 million (Farber, Cromar, and Davis 1999). The 2000 coverage measurement program calls for a sample of 5,000 small block clusters, much more than the number sampled in 1990.

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. Current plans for the 2000 census call for routinely searching a single ring of surrounding blocks, but conducting a “targeted extended search” that includes more distant blocks for a sample of clusters with large-scale errors. 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. Further analysis of the causes of these errors would be useful but was beyond the scope of our study.

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 phenomenon explains one of the problems that occurred in estimation from the 1990 PES. The estimated undercount for Asian men in the Northeast was extreme due to one sampled cluster with a huge undercount. The sample-based variance estimate was also quite large, so without “pre-smoothing” of variances, the undercount estimate was moderated by being strongly pulled in toward regression estimates under the empirical Bayes smoothing model. Model-based pre-smoothing of variances greatly reduced the sampling variance estimate for this poststratum, so there was relatively little smoothing of the undercount estimate. A further judgmental correction had to be made to pull this undercount estimate in to a more plausible value. We conjecture

that problems such as these in empirical Bayes inference, which follow from assuming normality of the data, might be solved by using hierarchical models with t error structure (Liu and Rubin 1998; Seltzer, Novak, and Lim 2000), extending our robust approach to a hierarchical structure.

5.3. More on Asymmetry

Our robust methods 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 application, the distribution of influence statistics might not be symmetrical, and this asymmetry might be systematically related to characteristics used in poststratification. For example, there may be large-scale errors, such as omission or duplication of an entire apartment building or trailer park, which are not due to misgeocoding and which therefore do not balance in expectation. These errors may tend to take place in certain types of geographical areas or areas with certain demographic characteristics, so the corresponding tail of the influence distribution is heavier in those areas.

Another possibly systematic relationship has to do with the effect of large weights. The estimated net undercount is the (weighted) difference between gross omissions and gross erroneous enumerations. Conditional on the underlying rates, this difference is distributed approximately as the difference of two scaled Poisson variables. In areas with relatively high rates of omissions relative to erroneous enumerations, this distribution has a heavy right tail; if some of these differences are inflated by large sampling weights, the distribution of the influence statistics inherits this heavy right tail. This argument suggests that downweighting of influential clusters will tend to downwardly bias undercount estimates in more heavily undercounted areas.

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; see Liepins (1983) for similar caveats with respect to automated data editing. 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 estimands

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 2 make this a promising area for continuing 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 has already been 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.
- Chernick, M., D. Downing, and D. Pike (1982). Detecting outliers in time-series data. *Journal of the American Statistical Association* 77, 743–747.
- Chernick, M. and V. Murthy (1983). The use of influence functions for outlier detection and data editing. *American Journal of Mathematical and Management Sciences* 3, 47–61.
- 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.
- Farber, J., R. Cromar, and P. Davis (1999). Accuracy and coverage evaluation survey: Initial listing sample results. Technical Report Procedures and Operations Memorandum Series R-16, June 25, 1999, U.S. Bureau of the Census, Decennial Statistical Studies Division.
- Fernandez, C. and M. Steel (1999). Multivariate student-t regression models: Pitfalls and inference. *Biometrika* 86, 153–167.

- Gwet, J. and L. Rivest (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association* 87, 1174–1182.
- Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69, 383–393.
- Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.
- Hidiroglou, M. and K. Srinath (1981). Some estimators of a population total containing large units. *Journal of the American Statistical Association* 78, 690–695.
- 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.
- Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology* 21, 79–87.
- Jaeckel, L. (1972). The infinitesimal jackknife. Technical report, Bell Laboratories, Murray Hill, New Jersey.
- Lange, K. and J. S. Sinsheimer (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* 2, 175–198.
- Lange, K. L., R. J. A. Little, and J. M. G. Taylor (1989). Robust statistical modeling using the t -distribution. *Journal of the American Statistical Association* 84, 881–896.
- 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.
- Liepins, G. (1983). Can automatic data editing be justified? One person’s opinion. In T. Wright (Ed.), *Statistical Methods and the Improvement of Data Quality*, pp. 205–213. Orlando, FL: Academic Press.

- Little, R. and P. Smith (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association* 82, 58–68.
- 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.
- Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.
- Mulry, M. and B. Spencer (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of the American Statistical Association* 88, 1080–1091.
- Potter, F. (1988). Survey of procedures to control extreme sampling weights. In *Proceedings of the ASA Section on Survey Research Methods*, pp. 453–458.
- 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.
- Seltzer, M., J. Novak, and N. Lim (2000). Robustifying hierarchical models for continuous response variables: The use of the t distribution in downweighting outlying observations. *Journal of Educational and Behavioral Statistics* 25, (in press).
- Smith, P. (1997). Winsorisation: Effects in practice. Technical Report Technical Report MQ-040, Office for National Statistics, London, England.
- Smith, P., P. Kokic, and S. Hibbitt (1997). Two-sided Winsorisation for the annual business inquiry (ABI). Technical Report MQ-045, Office for National Statistics, London, England.
- Stokes, L. (1990). A comparison of truncation and shrinking of sample weights. In *Proceedings of the 1990 Annual Research Conference*, Washington, DC, pp. 463–471. U.S. Bureau of the Census.
- West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B, Methodological* 46, 431–439.
- Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Table 1: Definitions of poststratification variables

Variable	Categories	Definition	<i>N</i>
Race	Black (not Hispanic)	% Black > 20	987
	Hispanic	% Hispanic > 10	953
	Other	(remainder)	3,350
Tenure	Owner	% owner > 50%	3,740
	Renter	% owner < 50%	1,550
Urbanicity	Urban	Center city > 250,000	2,275
	Suburban	Other urban area	1,827
	Rural	Not in urban area	1,188

ddadfasdfa

Table 2: Undercount estimates and estimated standard errors, by poststratum and nationally, for three estimation schemes: no downweighting (normal); downweighting based on the t distribution with $\nu = 2.5$ ($t_{2.5}$); and downweighting based on the Huber ψ -function with $c = 0.5$ (Huber_{0.5}). “SE%” gives the standard error of the robust estimator as a percentage of that of the conventional estimator. Sample size n is the number of blocks in the poststratum.

Poststratum	n	Normal		UC	$t_{2.5}$		Huber _{0.5}		
		UC	SE		SE%	UC	SE	SE%	
Black Rural Owner	60	5.61	1.55	3.50	1.03	66	3.74	0.87	56
Black Suburban Owner	137	1.19	1.45	1.78	0.71	49	1.89	0.78	54
Black Urban Owner	357	3.72	0.71	3.30	0.42	59	3.29	0.36	51
Black Rural Renter	6	11.75	2.95	9.58	1.04	35	10.00	1.27	43
Black Suburban Renter	116	0.95	2.63	4.53	0.92	35	4.25	0.83	32
Black Urban Renter	311	4.63	1.51	5.03	0.64	42	4.84	0.61	40
Hispanic Rural Owner	74	2.16	1.30	3.17	1.13	87	2.34	1.04	80
Hispanic Suburban Owner	163	2.64	0.89	2.15	0.56	63	2.24	0.63	71
Hispanic Urban Owner	295	2.41	0.62	2.67	0.38	61	2.58	0.37	60
Hispanic Rural Renter	11	8.02	2.62	5.51	5.01	191	5.74	5.21	199
Hispanic Suburban Renter	100	8.22	2.48	4.61	0.84	34	4.79	0.80	32
Hispanic Urban Renter	310	4.94	0.91	5.55	0.57	63	5.37	0.58	64
Other Rural Owner	932	0.43	0.42	0.76	0.18	43	0.73	0.17	40
Other Suburban Owner	948	1.12	0.32	0.54	0.13	41	0.60	0.12	38
Other Urban Owner	774	0.30	0.48	0.44	0.14	29	0.43	0.13	27
Other Rural Renter	105	8.64	5.28	5.96	1.27	24	5.70	1.33	25
Other Suburban Renter	363	1.98	1.77	2.29	0.40	23	2.45	0.39	22
Other Urban Renter	228	1.68	0.76	2.90	0.49	64	2.72	0.42	55
National	5290	1.76	0.22	1.65	0.09	41	1.59	0.09	41

Poststratum	<i>n</i>	Normal		<i>t</i> _{2.5}			UC
		UC	SE	UC	SE	SE%	
Black Rural Owner	60	5.61	1.55	3.50	1.03	66	3.0
Black Suburban Owner	137	1.19	1.45	1.78	0.71	49	1.3
Black Urban Owner	357	3.72	0.71	3.30	0.42	59	3.0
Black Rural Renter	6	11.75	2.95	9.58	1.04	35	10.0
Black Suburban Renter	116	0.95	2.63	4.53	0.92	35	4.0
Black Urban Renter	311	4.63	1.51	5.03	0.64	42	4.0
Hispanic Rural Owner	74	2.16	1.30	3.17	1.13	87	2.0
Hispanic Suburban Owner	163	2.64	0.89	2.15	0.56	63	2.0
Hispanic Urban Owner	295	2.41	0.62	2.67	0.38	61	2.0
Hispanic Rural Renter	11	8.02	2.62	5.51	5.01	191	5.0
Hispanic Suburban Renter	100	8.22	2.48	4.61	0.84	34	4.0
Hispanic Urban Renter	310	4.94	0.91	5.55	0.57	63	5.0
Other Rural Owner	932	0.43	0.42	0.76	0.18	43	0.0
Other Suburban Owner	948	1.12	0.32	0.54	0.13	41	0.0
Other Urban Owner	774	0.30	0.48	0.44	0.14	29	0.0
Other Rural Renter	105	8.64	5.28	5.96	1.27	24	5.0
Other Suburban Renter	363	1.98	1.77	2.29	0.40	23	2.0
Other Urban Renter	228	1.68	0.76	2.90	0.49	64	2.0
National	5290	1.76	0.22	1.65	0.09	41	1.0

Table 3: Undercount estimates and estimated standard errors by poststratum with no t -based downweighting and different ceilings for the sampling weights

Poststratum	No ceiling		Ceiling = 2000			Equal Weighting		
	UC	SE	UC	SE	SE%	UC	SE	SE%
Black Rural Owner	5.61	1.55	5.24	1.43	92	5.11	1.26	81
Black Suburban Owner	1.19	1.45	1.82	1.02	70	2.42	0.72	50
Black Urban Owner	3.72	0.71	3.84	0.68	96	3.64	0.56	79
Black Rural Renter	11.75	2.95	11.75	2.95	100	11.72	2.90	98
Black Suburban Renter	0.95	2.63	3.34	1.55	59	4.40	1.46	56
Black Urban Renter	4.63	1.51	4.89	1.09	72	5.52	0.77	51
Hispanic Rural Owner	2.16	1.30	2.51	1.60	123	4.15	1.47	113
Hispanic Suburban Owner	2.64	0.89	2.92	0.69	78	2.79	0.58	65
Hispanic Urban Owner	2.41	0.62	3.00	0.48	77	3.38	0.51	82
Hispanic Rural Renter	8.02	2.62	7.99	3.00	115	10.17	5.72	218
Hispanic Suburban Renter	8.22	2.48	7.58	2.22	90	6.13	1.06	43
Hispanic Urban Renter	4.94	0.91	5.14	0.69	76	6.42	0.70	77
Other Rural Owner	0.43	0.42	0.76	0.36	86	1.65	0.35	83
Other Suburban Owner	1.12	0.32	0.97	0.29	91	1.20	0.36	112
Other Urban Owner	0.30	0.48	0.32	0.37	77	0.32	0.30	62
Other Rural Renter	8.64	5.28	6.19	1.72	33	12.19	4.11	78
Other Suburban Renter	1.98	1.77	3.24	0.92	52	3.14	0.55	31
Other Urban Renter	1.68	0.76	1.16	0.96	126	-0.43	1.61	212
National	1.76	0.22	1.98	0.17	77	2.72	0.17	77

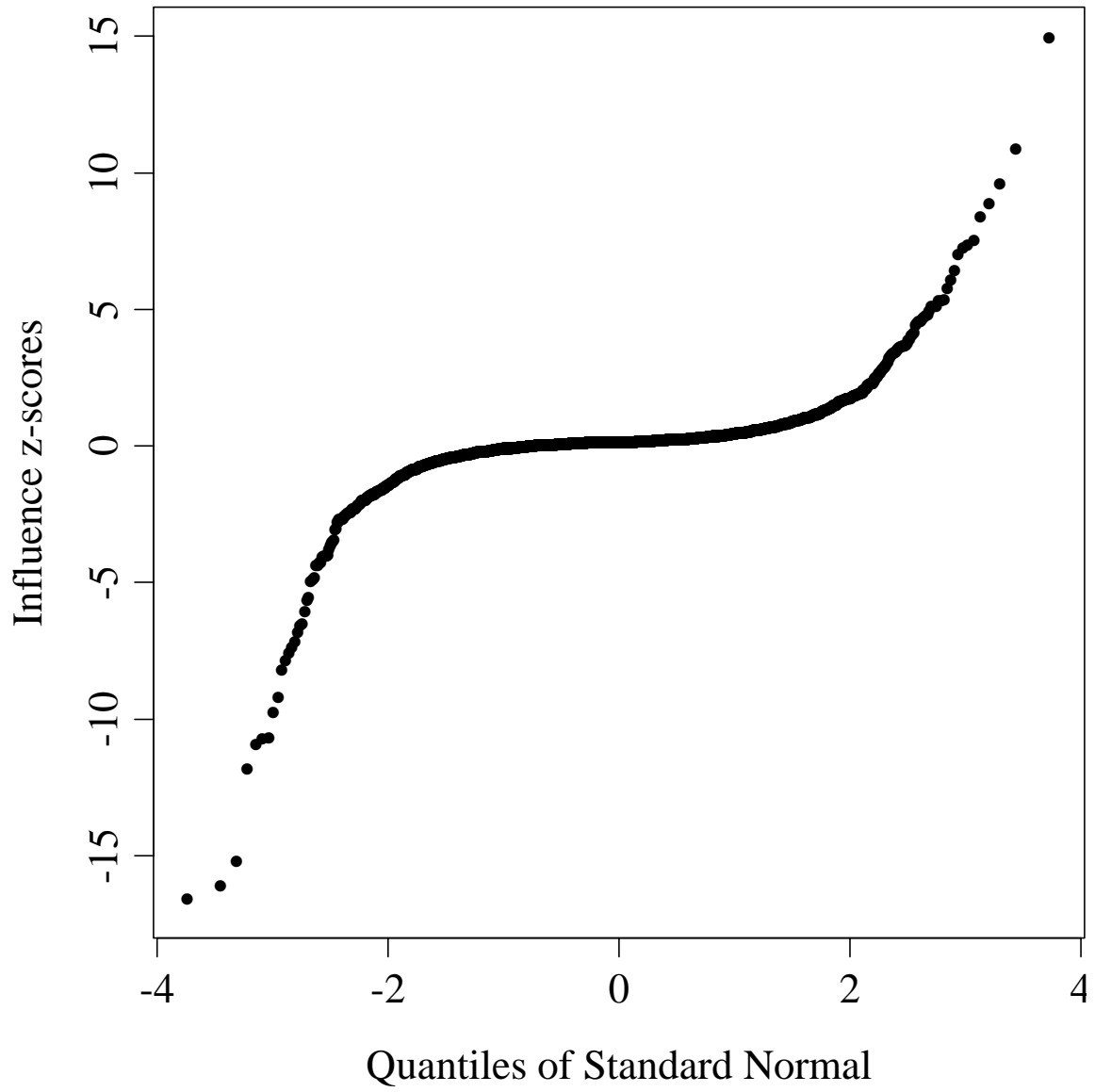


Figure 1: Normal quantile plot of influence z -scores

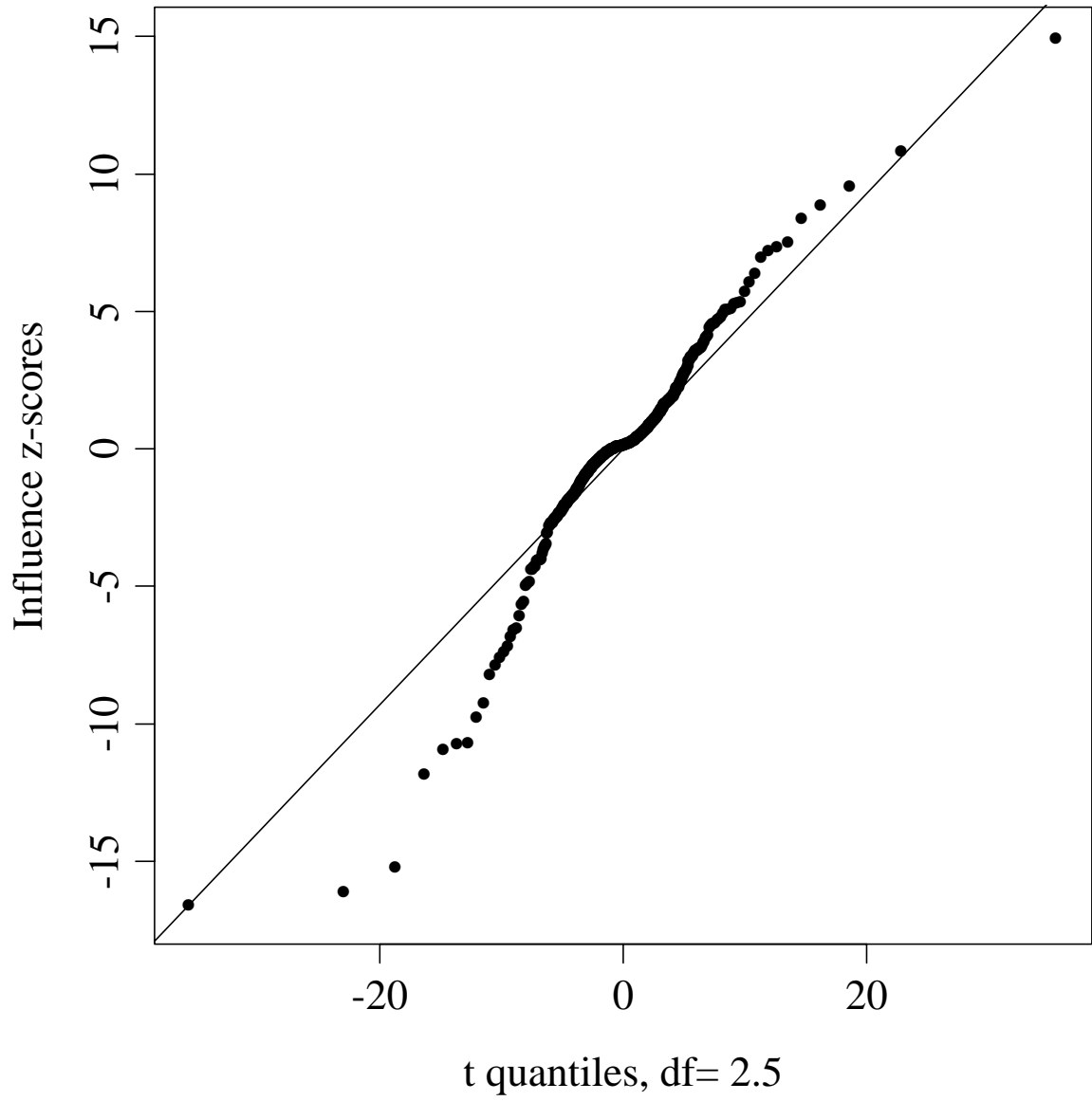


Figure 2: $t_{2.5}$ quantile plot of influence z -scores

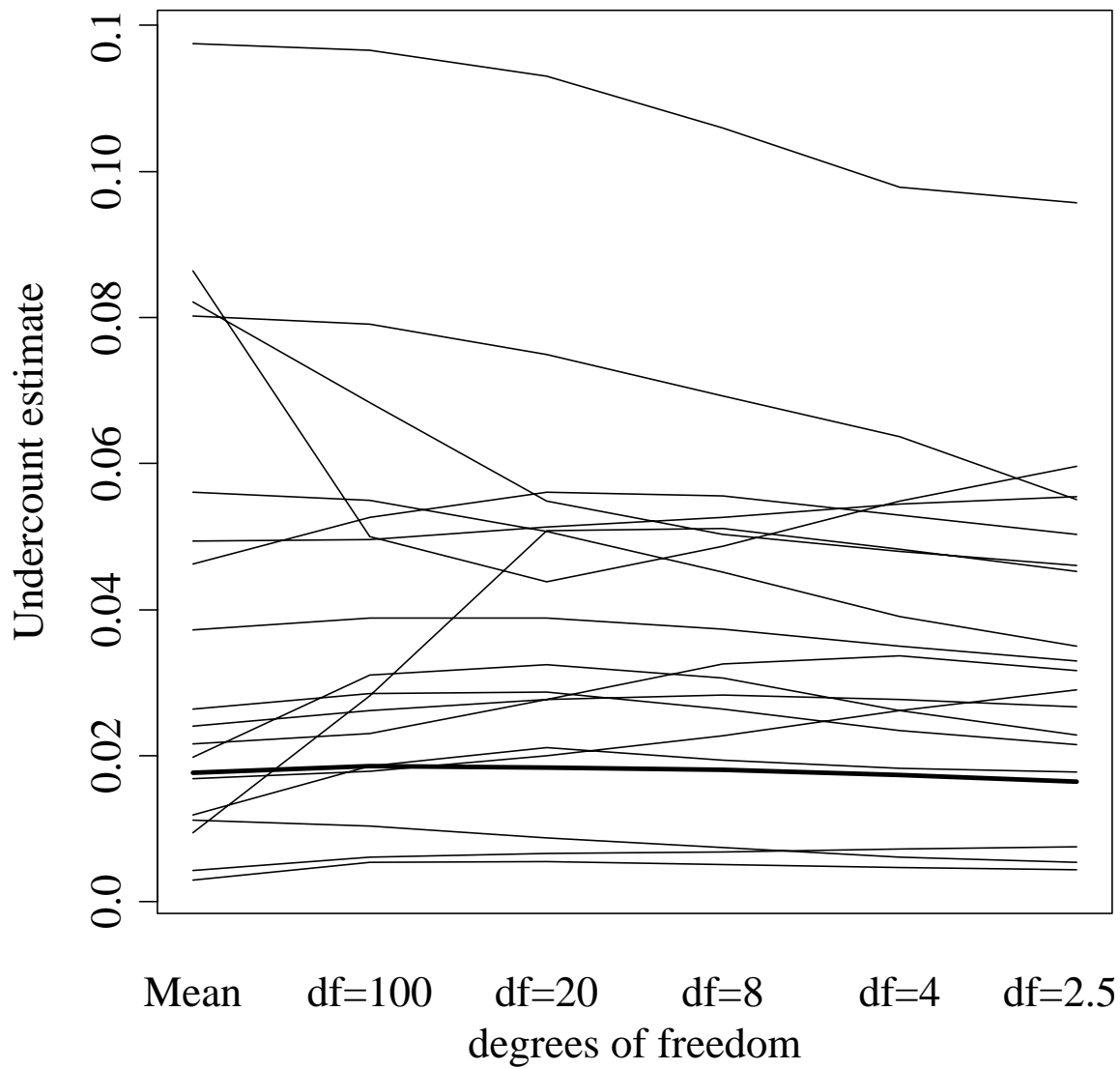


Figure 3: Traceplot of undercount estimates by poststratum and nationally (heavy line) against tuning parameter ν of robust estimator.

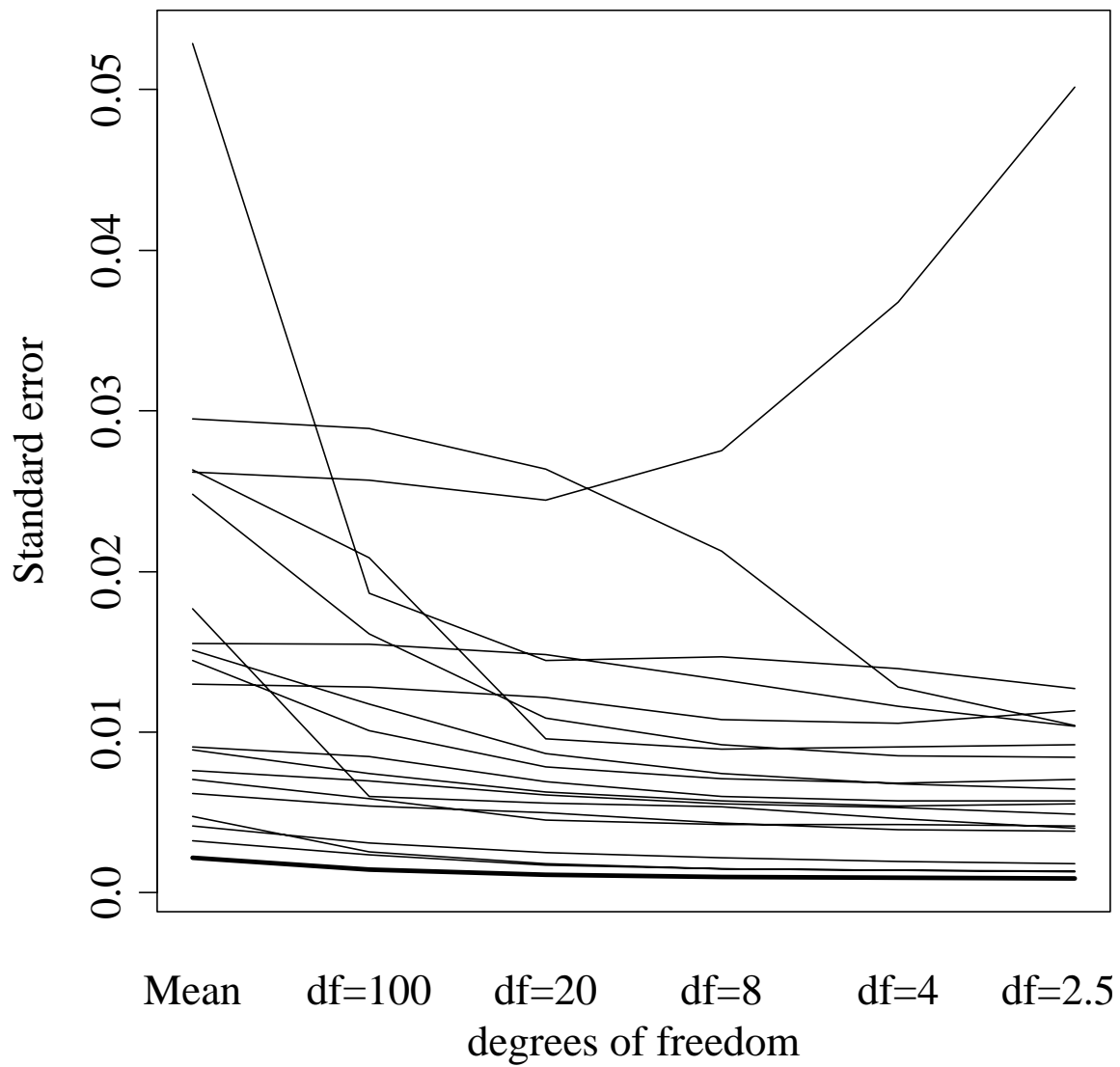


Figure 4: Traceplot of estimated standard errors by poststratum and nationally (heavy line) against tuning parameter ν of robust estimator.

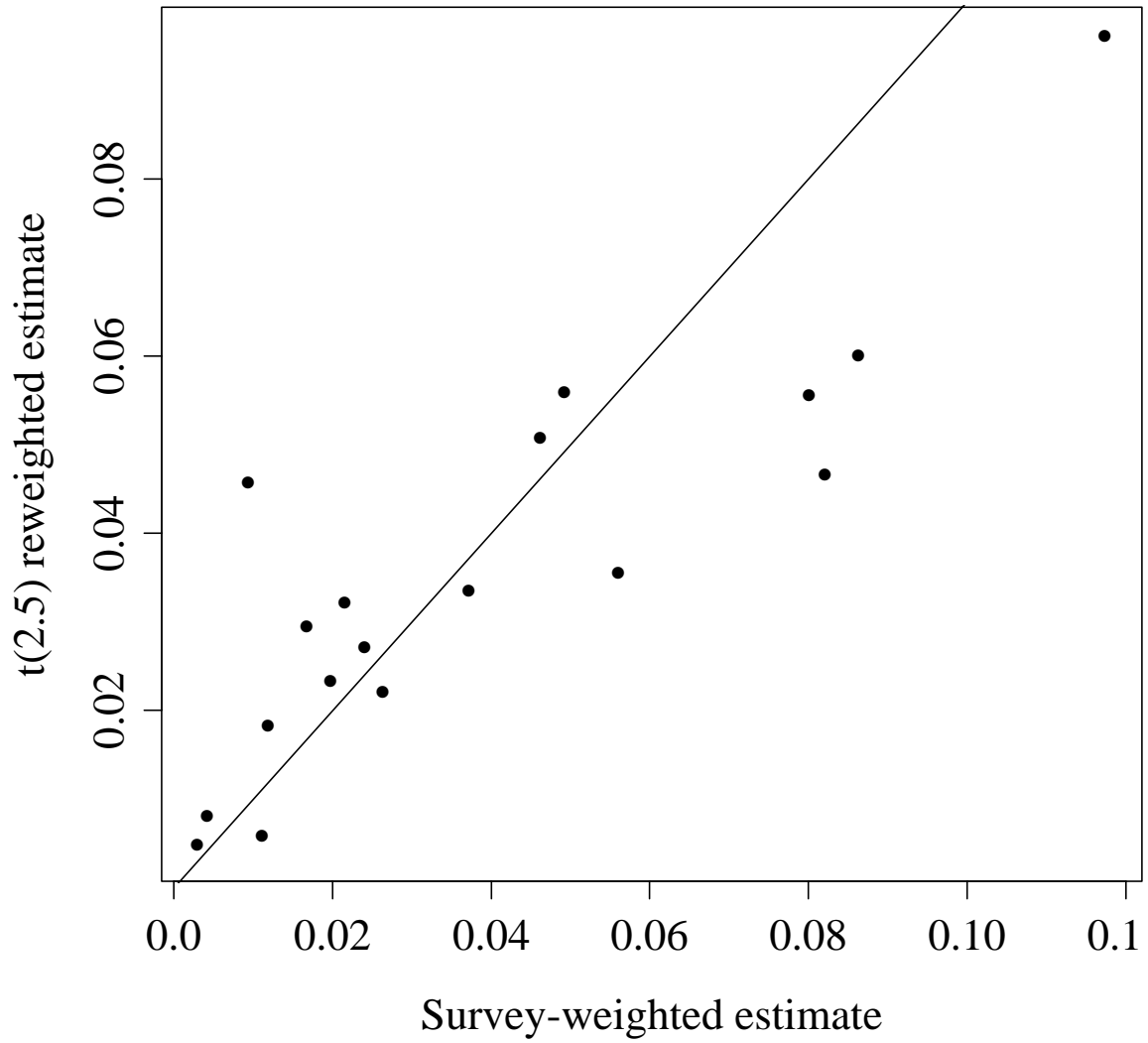


Figure 5: Undercount estimates by poststratum, without downweighting and with downweighting under $t_{2.5}$ procedure.

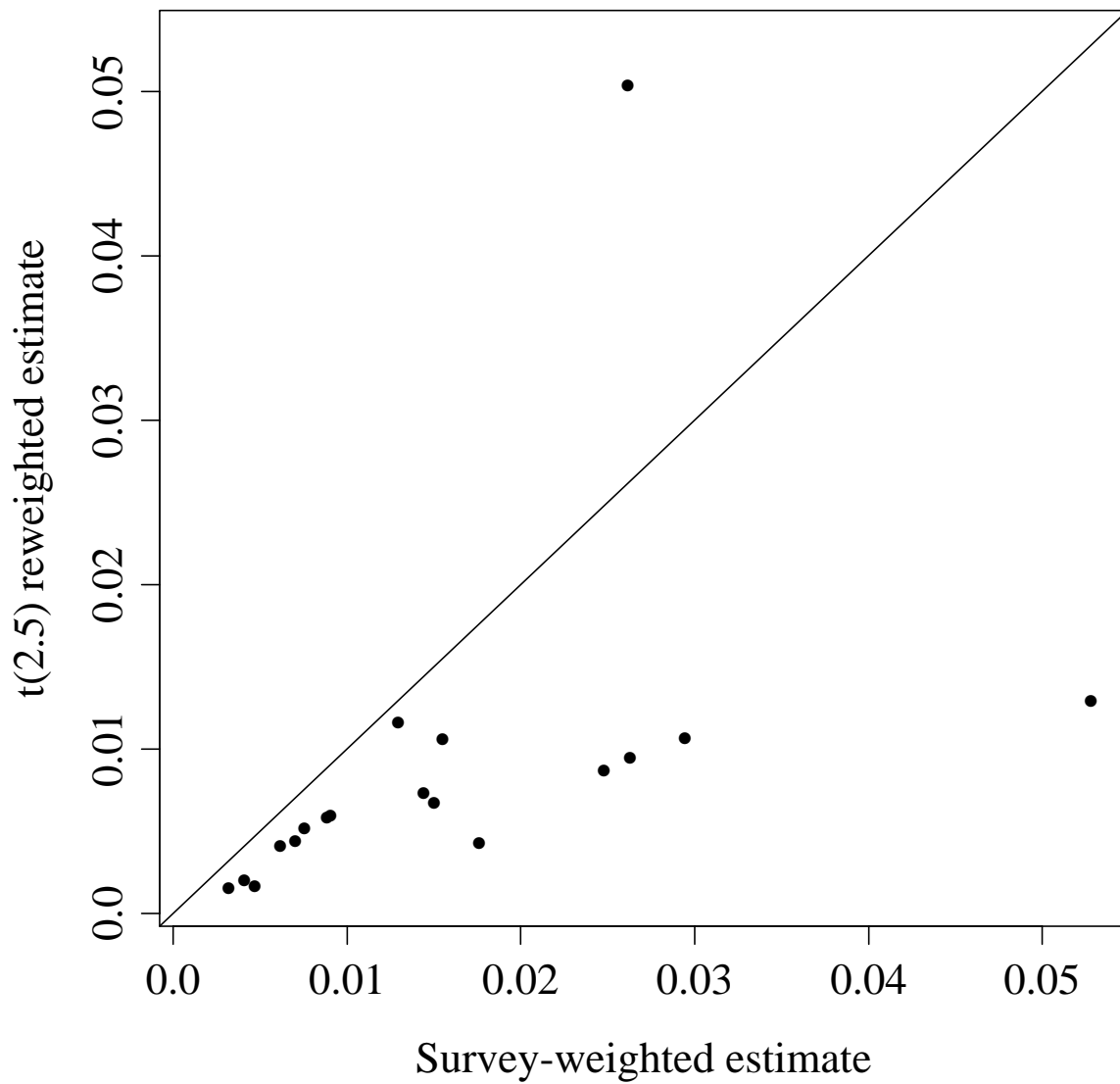


Figure 6: Estimated standard errors by poststratum, without downweighting and with downweighting under $t_{2.5}$ procedure.

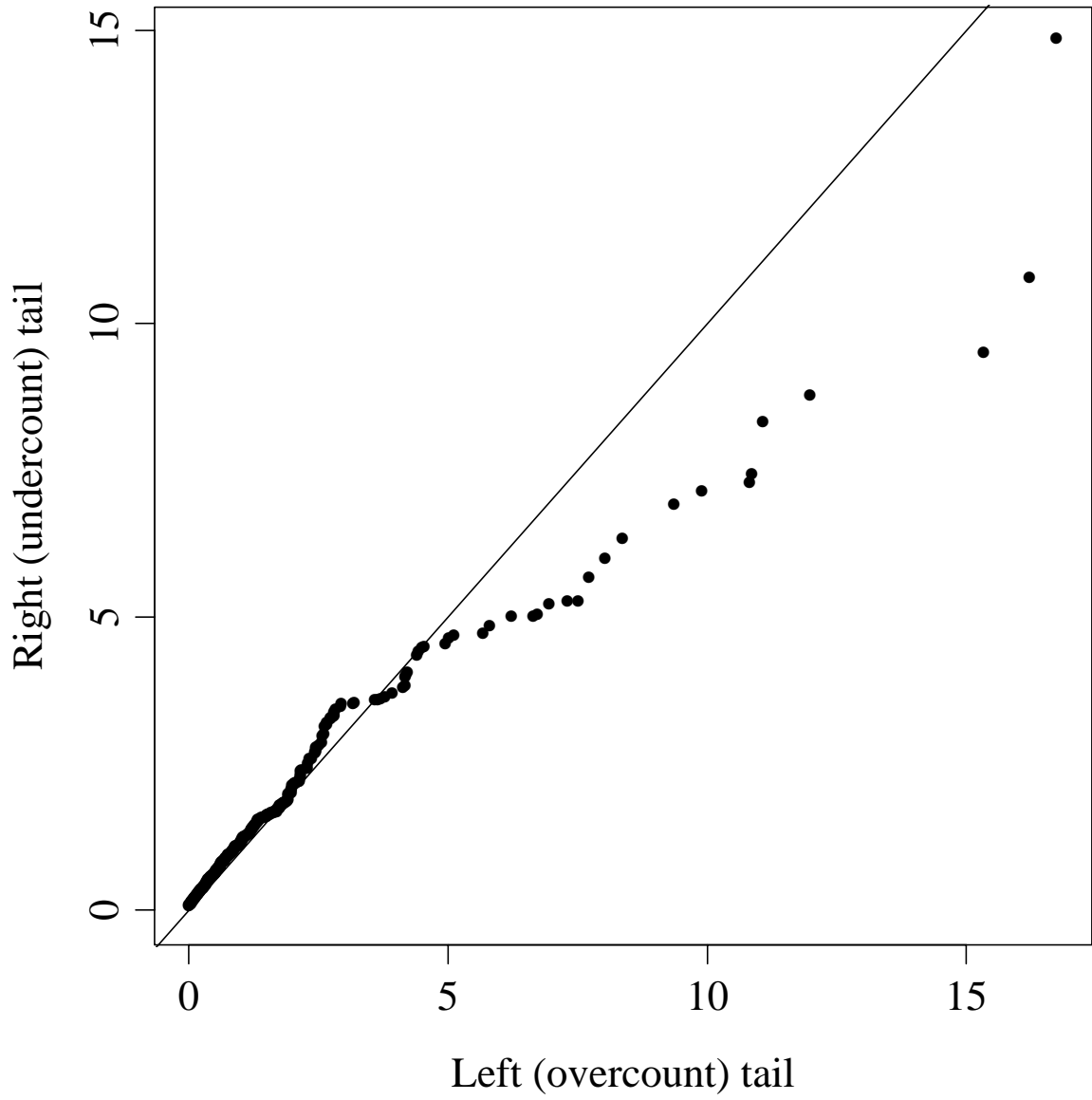


Figure 7: Quantile plot of left against right tail of influence z -score distribution.