DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 61–75, http://dx.doi.org/10.1515/JOS-2015-0003

Linear Regression Diagnostics in Cluster Samples

Jianzhu Li[1] *and Richard Valliant*[2]

An extensive set of diagnostics for linear regression models has been developed to handle nonsurvey data. The models and the sampling plans used for finite populations often entail stratification, clustering, and survey weights, which renders many of the standard diagnostics inappropriate. In this article we adapt some influence diagnostics that have been formulated for ordinary or weighted least squares for use with stratified, clustered survey data. The statistics considered here include DFBETAS, DFFITS, and Cook's D. The differences in the performance of ordinary least squares and survey-weighted diagnostics are compared using complex survey data where the values of weights, response variables, and covariates vary substantially.

Key words: Cook's D; DFBETAS; DFFITS; influence; model fitting; outlier; residuals.

1. Introduction

Linear regression models and estimators are often applied to analyze complex survey data using the pseudo maximum likelihood (PML) method (e.g., Binder 1983; Skinner et al. 1989).

A sample is considered to be informative when an unweighted model fitted to the sample data is different from the model fitted to the full population (Chambers and Skinner 2003). In such a case, using survey weights in PML estimation accounts for the informativeness. Using the sample weights in the regression estimator not only allows the analysts to account for the design features which govern the data collection process, but also provides a limited type of robustness to model misspecification (Pfeffermann and Holmes 1985; DuMouchel and Duncan 1983; Kott 1991). The sandwich estimator, the Taylor Series linearization estimator (Binder 1983; Fuller 2002), or some type of replication estimator (Wolter 2007) is often employed to obtain both design- and model-consistent variance estimators for the regression parameters. The analyses in this article cover the case in which survey weights are used in regression analysis. If the design is actually noninformative, the diagnostics developed here still apply even though the weights could, in principle, be omitted from model estimation.

Limited attention has been given to diagnosing the adequacy of working models and, more specifically, to detecting outlying and influential observations for regressions using

[1] Westat, 1600 Research Boulevard, Rockville MD 20850, USA. Email: JaneLi@westat.com
[2] Universities of Michigan and Maryland, 1218 Lefrak Hall, College Park MD 20742, USA. Email: rvallian@umd.edu

Acknowledgments: This article is based upon work partially supported by the National Science Foundation under Grant No. 0617081. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

complex survey data. Different threads of research cover locating and trimming extreme sample weights (Potter 1988, 1990), controlling the effect of outliers on the estimation of descriptive population statistics, and constructing outlier-robust estimation techniques (Chambers et al. 1993; Chambers 1996; Zaslavsky et al. 2001). Henry and Valliant (2012) review much of this literature. Diagnostics for regression models fitted from survey data are a more recent development. Korn and Graubard (1999) and Elliott (2007) introduced techniques for the evaluation of the quality of regressions on complex survey data. Li and Valliant (2009, 2011a, 2011b) examined leverages and methods of identifying influential single observations and groups of observations in single-stage samples. Liao and Valliant (2012a, 2012b) looked at condition indexes and variance inflation factors for linear regressions. In this article we will extend the work of Li and Valliant (2011a) for single-stage samples to samples that use stratification and clustering. We adapt the standard diagnostics – DFBETAS, DFFFITS, and Cook's D – to linear regression models fitted to clustered survey data.

Section 2 specifies the sample design we study, the model that will be used, and a variance estimator that is useful when developing diagnostics. Section 3 presents some diagnostics for identifying single observations that may be influential in fitting a model. Residuals, DFBETAS, DFFITS, and Cook's D are adapted for models fit using stratified, clustered data. In the fourth section, the new diagnostics are illustrated using a data set taken from a large U.S. household survey. Section 5 forms the conclusion.

2. Model Specification and Variance Estimation

To formulate regression diagnostics for clustered survey data, models will be used. Suppose the population contains $h = 1, \ldots, H$ strata, $i = 1, \ldots, N_h$ clusters in stratum h, and $k = 1, \ldots, M_i$ units in cluster hi. A two-stage stratified sample of units is selected with n_h clusters or primary sampling units (PSUs) sampled at the first stage in stratum h with replacement (although without-replacement is more common in practice, a with-replacement formulation has the advantage of producing simpler design-based variance formulas that are more informative for the analyses in this article). The total number of sample clusters is $n = \sum_{h=1}^{H} n_h$. Let m_{hi} be the number of sampled units in the (hi)th cluster, $m = \sum_{h=1}^{H} \sum_{i \in s_h} m_{hi}$, with s_h being the sample of clusters in stratum h, and w_{hik} be the sample weight of the kth unit in the (hi)th cluster. The average number of sample units per sample cluster is $\bar{m} = m/n$. Suppose that $\mathbf{x}_{hik}$ is a p-vector of explanatory variables for unit k in cluster hi and that a variable Y_{hik} collected in the survey follows the linear model:

$$
\begin{aligned}
Y_{hik} &= \mathbf{x}_{hik}^T \boldsymbol{\beta} + \varepsilon_{hik} \\
Cov_M(\varepsilon_{hik}, \varepsilon_{h'i'k'}) &= \begin{cases} \sigma^2 & h = h', i = i', k = k' \\ \sigma^2 \rho & h = h', i = i', k \neq k' \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{1}
$$

This model posits that all units have a common variance and the intracluster correlation, ρ, is the same for all clusters. Units in different clusters are uncorrelated. In practice, ρ is

usually positive and can be estimated using analysis of variance (ANOVA) or related methods. The survey-weighted (SW) estimator of $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}}_{SW} = \sum_{h=1}^{H} \sum_{i \in s_h} \sum_{k \in s_{hi}} \mathbf{A}^{-1} \mathbf{x}_{hik} w_{hik} Y_{hik} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{Y}_{hi}$$

with s_{hi} being the sample of units from cluster hi, and

$\mathbf{X}_{hi}$ = the $m_{hi} \times p$ matrix of the $\mathbf{x}_{hik}$ for the m_{hi} sample units in cluster hi;
$\mathbf{W}_{hi}$= the $m_{hi} \times m_{hi}$ diagonal matrix of survey weights for sample units in sample cluster hi;
$\mathbf{Y}_{hi}$ = the m_{hi}-vector of Y_{hik}'s for sample units in cluster hi, and

$$\mathbf{A} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{X}_{hi}.$$

For later use we also define $\mathbf{X}_h^T = (\mathbf{X}_{h1}^T, \ldots, \mathbf{X}_{hn_h}^T)$, $\mathbf{X}^T = (\mathbf{X}_1^T, \ldots, \mathbf{X}_H^T)$, and $\mathbf{W}_h = \text{blkdiag}(\mathbf{W}_{hi})_{i \in s_h}$. Under (1) the model variance of $\hat{\boldsymbol{\beta}}_{SW}$ is

$$\begin{aligned} \text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right) &= \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \, \text{var}_M(\mathbf{Y}_{hi}) \mathbf{W}_{hi} \mathbf{X}_{hi} \mathbf{A}^{-1} \qquad (2) \\ &= \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \left((1-\rho)\sigma^2 \mathbf{I}_{m_{hi}} + \rho\sigma^2 \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \right) \mathbf{W}_{hi} \mathbf{X}_{hi} \mathbf{A}^{-1} \end{aligned}$$

where $\mathbf{I}_{m_{hi}}$ is the $m_{hi} \times m_{hi}$ identity matrix and $\mathbf{1}_{m_{hi}}$ is a vector of m_{hi} 1s. To test the significance of $\hat{\boldsymbol{\beta}}_{SW}$ or its components, the sandwich estimator in Binder (1983) or the linearization estimator in Fuller (2002) is typically used. Both of these have design-based and model-based justifications. In fact, the sandwich estimator is approximately model unbiased under a model more general than (1), in which the errors are correlated within each cluster but the particular form of the correlation is unspecified (e.g., see Valliant et al. 2000, chap. 9). However, to motivate cutoff values for identifying extremes based on the diagnostics in Section 3, the form of the variance in (2) is useful. Estimates of the components of (2) are needed, and a workable approach is to use purely model-based estimators. To that end, define $\hat{\boldsymbol{\beta}}_{OLS} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}_{OLS}^{-1} \mathbf{X}_{hi}^T \mathbf{Y}_{hi}$ with $\mathbf{A}_{OLS} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{X}_{hi}$ to be the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$, and $e_{hik} = Y_{hik} - \mathbf{x}_{hik}^T \hat{\boldsymbol{\beta}}_{OLS}$ to be the residual calculated from the OLS estimator. Using these residuals, define

$$\hat{P} = \frac{1}{n} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{1}{m_{hi} - 1} \sum_{k \in s_{hi}} (e_{hik} - \bar{e}_{hi})^2$$

$$\hat{Q} = \sum_{h=1}^{H} \sum_{i \in s_h} m_{hi} (\bar{e}_{hi} - \bar{e}_h)^2 \Big/ (n-1)$$

$$\hat{D} = \left(m - \sum_h \sum_{i \in s_h} m_{hi}^2 \Big/ m \right) \Big/ (n-1),$$

where $\bar{e}_{hi} = \sum_{k \in s_{hi}} e_{hik}/m_{hi}$ and $\bar{e}_h = \sum_{i \in s_h} \sum_{k \in s_{hi}} e_{hik} / \sum_{i \in s_h} m_{hi}$. Using $\hat{P}$, $\hat{Q}$, and $\hat{D}$, we can formulate estimators as:

$$\begin{aligned} \widehat{(1-\rho)\sigma^2} &= \hat{P} \\ \widehat{\rho\sigma^2} &= (\hat{Q} - \hat{P})/\hat{D} \end{aligned} \tag{3}$$

These are similar to the estimators in Valliant et al. (2000, sec. 8.3.1) for a common-mean model. Showing that they are model-unbiased for $\rho\sigma^2$ and $(1-\rho)\sigma^2$ is straightforward. Another alternative is to use ANOVA or restricted maximum-likelihood methods in, for instance, SAS® `proc varcomp` or `proc mixed` or Stata® `xtmixed` or the `lmer` function in the R package `lme4` (Bates et al. 2012).

When $\widehat{\rho\sigma^2}$ and $\widehat{(1-\rho)\sigma^2}$ are available, the estimated variance of $\hat{\boldsymbol{\beta}}$ under Model (1) can be constructed as

$$v_M(\hat{\boldsymbol{\beta}}_{SW}) = \sum_h \sum_{s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \left(\widehat{(1-\rho)\sigma^2} \mathbf{I}_{m_{hi}} + \widehat{\rho\sigma^2} \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \right) \mathbf{W}_{hi} \mathbf{X}_{hi} \mathbf{A}^{-1} \tag{4}$$

This variance estimator is highly dependent on the working model and is not robust to departures from that model. Because of its nonrobustness, a sandwich or replication estimator is preferred for actually estimating the variance of $\hat{\boldsymbol{\beta}}_{SW}$. However, (4) does have some advantages in determining cutoffs for diagnostics, as described subsequently.

There are alternatives to the estimators of $\rho\sigma^2$ and $(1-\rho)\sigma^2$ in (3). Pfeffermann et al. (1998) proposed the probability-weighted iterative generalized least squares (PWIGLS) estimator to obtain consistent estimates of the population variance parameters σ_U^2 and ρ_U, i.e., the parameters that would be estimated from a census. The PWIGLS estimator, which assumes that the sampling probabilities for both stages π_{hi} and $\pi_{k|hi}$, or equivalently their inverses, w_{hi} and $w_{k|hi}$, are known, is adapted from the standard iterative generalized least squares procedure by analogy with PML. Alternative inflation-type estimators using the two-level sample weights have also been considered (Longford 1995; Graubard and Korn 1996). However, Korn and Graubard (2003) later showed that these estimators can be severely biased when the sampling is informative. They proposed a new set of estimators for variance components that would be approximately unbiased regardless of the sampling design. The limitation of these estimators is that they require knowledge of the second-order inclusion probabilities of the observations. In many surveys, analysts will not know the value of w_{hi}, $w_{k|hi}$, or the joint inclusion probabilities. Consequently, we use the estimators in (3) which are always feasible.

3. Identifying Single Influential Observations

The diagnostic tools presented here are designed to measure the discrepancy in estimated regression coefficients and fitted values, between fitting linear models with and without potentially influential points.

3.1. Residuals

Residuals, which can be used to filter points with outlying Y values, usually are standardized to have unit model variance. For clustered sampling and its corresponding

model (1), we can divide e_{hik} by $\hat{\sigma} = \sqrt{\hat{P} + (\hat{Q} - \hat{P})\hat{D}^{-1}}$; see (3). Generally, the standardized residuals are referred to the standard normal distribution to identify extreme points. If the e_{hik} are not normal, the Gauss inequality (Pukelsheim 1994) is useful for setting a cutoff value.

Gauss Inequality: If the distribution of a random variable X has a single mode at μ_0, then $P\{|X - \mu_0| > r\} \leq 4\tau^2/9r^2$ for all $r \geq \sqrt{4/3}\,\tau$, where $\tau^2 = E\left[(X - \mu_0)^2\right]$.

Suppose that under Model (1), in addition to having a mean of 0, the residuals have a mode of zero. Based on the Gauss Inequality with $r = 2\sigma$, the absolute value of a residual has a probability of about 90% of being less than twice its standard deviation, and with $r = 3\sigma$, it has a probability of about 95% of being less than three times its standard deviation. If we rescale the residuals by a consistent estimate $\hat{\sigma}$ of σ, either $r/\hat{\sigma} = 2$ or 3 can be used to identify outlying residuals, depending on an analyst's preference.

3.2. *DFBETAS*

The standard DFBETAS statistic (Belsley et al. 1980) measures the change in the estimate of $\boldsymbol{\beta}$ when a single unit is removed from the sample. The statistic is also standardized so that it can be referred to a standard normal distribution to determine which values are extreme enough to deserve scrutiny. First, note that (2) can be written as

$$\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right) = \sigma^2 \sum_{h=1}^{H} \sum_{s_h} \mathbf{C}_{hi}\mathbf{R}_{hi}\mathbf{C}_{hi}^T \tag{5}$$

where $\mathbf{R}_{hi} = \left[(1-\rho)\mathbf{I}_{m_{hi}} + \rho \mathbf{1}_{m_{hi}}\mathbf{1}_{m_{hi}}^T\right]$ and $\mathbf{C}_{hi} = \mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}$ with (jk)th element $c_{j,hik}$ $(j = 1, \ldots, p; k = 1, \ldots, m_{hi})$. The correlation ρ could be estimated as $\hat{\rho} = \left[1 + \hat{P}\hat{D}/(\hat{Q} - \hat{P})\right]$ or by some other model-based alternative. The variance estimator is then

$$v_M(\hat{\beta}_{SWj}) = \sigma^2 \sum_h \sum_{s_h} \left(c_{j,hi1} \ldots c_{j,him_{hi}}\right) \begin{pmatrix} 1 & & \hat{\rho} \\ & \ddots & \\ \hat{\rho} & & 1 \end{pmatrix} \left(c_{j,hi1} \ldots c_{j,him_{hi}}\right)^T$$

$$= \sigma^2 \sum_h \sum_{s_h} \left(\sum_{k=1}^{m_{hi}} c_{j,hik}^2 + \hat{\rho} \sum_{k \neq k'}^{m_{hi}} c_{j,hik}c_{j,hik'} \right).$$

To measure the difference in each estimated coefficient after the (hik)th unit is deleted, we define $\hat{\boldsymbol{\beta}}_{SW}(hik)$ as the parameter estimate after deleting unit k in cluster hi. The difference between the full sample estimate and the delete-one estimate, $\hat{\boldsymbol{\beta}}_{SW}(hik)$, can be found as

$$DFBETA_{hik} = \hat{\boldsymbol{\beta}}_{SW} - \hat{\boldsymbol{\beta}}_{SW}(hik) = \frac{\mathbf{A}^{-1}\mathbf{x}_{hik}e_{hik}w_{hik}}{1 - \tilde{h}_{hik,hik}},$$

where $\tilde{h}_{hik,hik} = \mathbf{x}_{hik}^T\mathbf{A}^{-1}\mathbf{x}_{hik}w_{hik}$ is the leverage of the (hik)th unit, which is the kth diagonal element of the matrix $\mathbf{H}_{hii} = \mathbf{X}_{hi}\mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}$ (see, e.g., Miller 1974;

Valliant et al. 2000, sec. 9.5). The DFBETAS statistic, which is standardized, is constructed as

$$DFBETAS_{hik,j} = \frac{c_{j,hik} e_{hik} / \left(1 - \tilde{h}_{hik,hik}\right)}{\sqrt{\text{var}_M(\hat{\beta}_{SWj})}} \tag{6}$$

$$= \frac{c_{j,hik}}{\sqrt{\sum_{s_h}\left(\sum_{k=1}^{m_i} c_{j,hik}^2 + \rho \sum_{k \neq l}^{m_i} c_{j,hik} c_{j,hil}\right)}} \cdot \frac{e_{hik}}{\sigma} \cdot \frac{1}{1 - \tilde{h}_{hik,hik}}.$$

Note that for actual calculations, a more robust sandwich or replication estimator of $\text{var}_M\left(\hat{\beta}_{SWj}\right)$ would be used in the denominator of (6). Using the diagonal element of (5) in the denominator of $DFBETAS_{hik,j}$ allows us to motivate a heuristic cutoff for identifying extremes.

In order to define a cutoff, some simplifications are needed. If the population and sample sizes from each cluster are bounded by $\bar{M}$ and $\bar{m}$, then $w_{hik} = O\left(N/n\right)$. If the xs are bounded, $\mathbf{C}_{hi} = O\left(n^{-1}\right)$ elementwise and the first term of (6) has order $n^{-1/2}$. Under the same conditions, $\tilde{h}_{hik,hik} = O\left(n^{-1}\right)$, and a rough cutoff after applying the Gauss inequality to e_{hik} would be $2/\sqrt{n}$ or $3/\sqrt{n}$.

A slightly more fine-tuned cutoff is obtained as follows. Following the developments in Scott and Holt (1982) as extended by Liao and Valliant (2012b), the model variance of $\hat{\boldsymbol{\beta}}_{SW}$ can be written as

$$\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right) = \sigma^2 \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{G}$$

where $\mathbf{G} = \left[\sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi}\right] \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1}$. The matrix $\mathbf{G}$ is a generalized design effect that measures the factor by which the model variance differs from that of weighted least squares when all units are uncorrelated. Under Model (1), we have

$$\sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi} = \sigma^2 \left[(1 - \rho) \mathbf{X}_h^T \mathbf{W}_h^2 \mathbf{X}_h + \rho \sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{W}_{hi}^2 \mathbf{X}_{Bhi}\right].$$

where $\mathbf{X}_{Bhi} = m_{hi}^{-1} \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \mathbf{X}_{hi}$ with $\mathbf{1}_{m_{hi}}$ being a vector of m_{hi} 1s. If the sample is self-weighting so that $w_{hik} \equiv w$, then under Model (1) $\mathbf{G}$ can be written as

$$\mathbf{G} = w\sigma^2 \left[\mathbf{I}_p + (\mathbf{M} - \mathbf{I}_p)\rho\right]$$

where $\mathbf{M} = \left(\sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{X}_{Bhi}\right) (\mathbf{X}^T \mathbf{X})^{-1}$ and $\mathbf{I}_p$ is the $p \times p$ identity matrix. If we assume that the sample size within every cluster is $m_{hi} = \bar{m}$ and that the vector of covariates for every element in cluster hi is the same, $\mathbf{x}_{hik} = \bar{\mathbf{x}}_{hi}$, with some algebra it

follows that

$$\sum_{i\in s_h} m_{hi}\mathbf{X}_{Bhi}^T\mathbf{X}_{Bhi} = \bar{m}\sum_h\sum_{s_h}\bar{\mathbf{x}}_{hi}\bar{\mathbf{x}}_{hi}^T$$

$$\mathbf{X}^T\mathbf{X} = \sum_h\sum_{s_h}\bar{\mathbf{x}}_{hi}\bar{\mathbf{x}}_{hi}^T.$$

Using these results, $\mathbf{M}$ reduces to $\bar{m}\mathbf{I}_p$. In these special circumstances, the model variance of the survey-weighted least squares estimator is

$$\text{var}_M(\hat{\boldsymbol{\beta}}_{SW}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\left[\mathbf{I}_p + \rho\times diag(\bar{m}-1)\mathbf{I}_p\right].$$

The model variance of the jth coefficient of $\hat{\boldsymbol{\beta}}_{SW}$, which is needed for $DFBETAS_{hik,j}$, is then

$$\text{var}_M(\hat{\beta}_{SWj}) = \sigma^2(\mathbf{X}^T\mathbf{X})_{jj}^{-1}[1+(\bar{m}-1)\rho]$$

where $(\mathbf{X}^T\mathbf{X})_{jj}^{-1}$ denotes the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. Assuming the xs are all bounded, the order of magnitude of each element of $(\mathbf{X}^T\mathbf{X})^{-1}$ is n^{-1}. Thus $\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SWj}\right) = O\left(n^{-1}\right)\left[1+(\bar{m}-1)\rho\right]$. Using $c_{j,hik} = O\left(n^{-1}\right)$, the first term in (6) is $c_{j,hik}\Big/\sqrt{\text{var}_M(\hat{\beta}_j)} \approx \left\{O(n)\left[1+(\bar{m}-1)\rho\right]\right\}^{-1/2}$. As a result, a somewhat more refined cutoff value for $DFBETAS_{ik,j}$ is $2\Big/\sqrt{n\left[1+(\bar{m}-1)\rho\right]}$ or $3\Big/\sqrt{n\left[1+(\bar{m}-1)\rho\right]}$.

3.3. DFFITS

Multiplying the DFBETA statistic by the $\mathbf{x}_{hik}^T$ vector, we obtain the measure of change in the (*hik)*th fitted values due to the deletion of the *(hik)*th observation,

$$DFFIT_{hik} = \hat{Y}_{hik} - \hat{Y}_{hik}(hik) = \frac{\tilde{h}_{hik,hik}e_{hik}}{1-\tilde{h}_{hik,hik}}.$$

The variance of the predicted value is

$$\text{var}_M\left(\hat{Y}_{hik}\right) = \mathbf{x}_{hik}^T\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right)\mathbf{x}_{hik}$$

$$= \sigma^2\sum_{i'\in s}\left(\sum_{k'=1}^{m_{hi'}}\tilde{h}_{hik,hi'k'}^2 + \rho\sum_{k''\neq k'}^{m_{hi'}}\tilde{h}_{hik,hi'k'}\tilde{h}_{hik,hi'k''}\right).$$

The DFFITS statistic is formulated as

$$DFFITS_{hik} = \frac{\tilde{h}_{hik,hik}e_{hik}/\left(1-\tilde{h}_{hik,hik}\right)}{\sqrt{\text{var}_M\left(\hat{Y}_{hik}\right)}}$$

We can make approximations analogous to the ones used for DFBETAS in order to justify a cutoff for DFFITS. Based on (7) for the special case of $m_{hi} = \bar{m}$ and $\mathbf{x}_{hik} = \bar{\mathbf{x}}_{hi}$, we have $v_M\left(\hat{Y}_{ik}\right) = \mathbf{x}_{ik}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left[\mathbf{I}_p + diag(\bar{m}-1)\rho\right]\mathbf{x}_{ik}$. Each element of $\mathbf{X}^T\mathbf{X}$ is the sum of m elements, and, if each x is bounded, is $O(m)$. The variance $\text{var}_M\left(\hat{Y}_{ik}\right)$ is a sum of

p elements; thus $v_M\left(\hat{Y}_{ik}\right) = O(p/m)\left[1 + (\bar{m} - 1)\rho\right]$. Since the average leverage is p/m, a rough value on $\frac{\tilde{h}_{hik,hik}/(1-\tilde{h}_{hik,hik})}{\sqrt{\text{var}_M(\hat{Y}_{hik})}}$ is $\frac{p/m}{1-p/m} \Big/ \sqrt{\frac{p}{m}\left[1 + (\bar{m} - 1)\rho\right]} = \sqrt{p/\left\{n\bar{m}\left[1 + (\bar{m} - 1)\rho\right]\right\}}$, assuming that the number of sample units, m, is much larger than the number of regressors, p. Thus a heuristic cutoff for the DFFITS statistic is $k\sqrt{p/n\bar{m}\left[1 + \hat{\rho}(\bar{m} - 1)\right]}$ with k being 2 or 3.

3.4. Modified Cook's Distance

Under the working Model (1), a quadratic statistic that measures the effect on the entire $\hat{\boldsymbol{\beta}}_{SW}$ vector of dropping the kth element in cluster hi can be constructed as

$$ED_{hik} = \left[\hat{\boldsymbol{\beta}}_{SW} - \hat{\boldsymbol{\beta}}_{SW}(hik)\right]^T \left[\text{var}\left(\hat{\boldsymbol{\beta}}_{SW}\right)\right]^{-1} \left[\hat{\boldsymbol{\beta}}_{SW} - \hat{\boldsymbol{\beta}}_{SW}(hik)\right]$$

where $\hat{\boldsymbol{\beta}}_{SW}(hik)$ is the parameter estimate after deleting unit k in cluster hi and $\text{var}\left(\hat{\boldsymbol{\beta}}_{SW}\right)$ is any of the variance estimators discussed in Section 1. To determine a heuristic cutoff value for ED_{ik}, we use the model variance $\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right)$ under (1) and write the statistic as

$$ED_{hik} = \left(\frac{e_{hik}}{\sigma}\right)^2 \frac{1}{\left(1 - \tilde{h}_{hik,hik}\right)^2} w_{hik}\mathbf{x}_{hik}^T \left[\mathbf{X}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{X}\right]^{-1} \mathbf{x}_{hik} w_{hik}$$

where the matrix $\mathbf{R}$ is block diagonal with 1 on the diagonal and ρ off the diagonal in each block (cluster); the dimension of block hi is $m_{hi} \times m_{hi}$. If the number of units within each sampled PSU, m_{hi}, is bounded, $w_{hik}\mathbf{x}_{hik}^T \left[\mathbf{X}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{X}\right]^{-1} \mathbf{x}_{hik} w_{hik} = O\left(n^{-1}\right)$, and using similar reasoning to that employed in Subsections 3.1 and 3.2, we arrive at a rough value for ED_{hik} of $p\left[n\bar{m}\left(1 + \hat{\rho}(\bar{m} - 1)\right)\right]^{-1}$. Therefore, in the clustered sampling case we can compare $\sqrt{ED_{hik}}$ with the cutoff value $2\sqrt{p/n\bar{m}\left[1 + \hat{\rho}(\bar{m} - 1)\right]}$ or $3\sqrt{p/n\bar{m}\left[1 + \hat{\rho}(\bar{m} - 1)\right]}$. A more convenient form is found by standardizing ED_{hik} and taking its square root. Based on the classic Cook's Distance, we term this the Modified Cook's distance:

$$MD_{hik} = \sqrt{\left\{n\bar{m}\left[1 + \hat{\rho}(\bar{m} - 1)\right]\right\} ED_{hik}/p}$$

and compare MD_{hik} to 2 or 3.

Table 1. Quantiles of variables in NHANES regression of systolic blood pressure on age, BMI, and blood lead

	Quantiles				
Variables	0%	25%	50%	75%	100%
Systolic BP	82	102	108	114	146
Age	20	22	24	27	29
BMI	14.42	22.84	26.43	31.62	61.68
Log(Lead+1)	0.18	0.47	0.64	0.83	3.75
Survey Weight	698.39	3,576.69	11,467.06	31,094.18	103,831.17

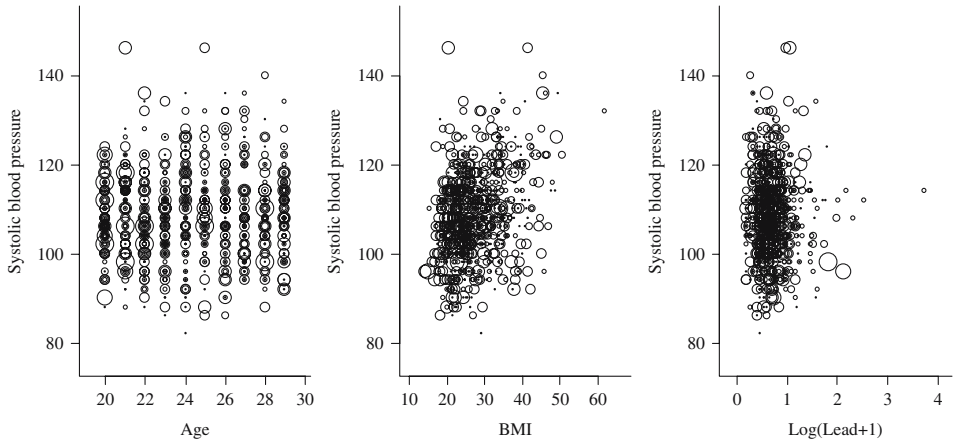

Fig. 1. Bubble plots of systolic blood pressure versus three auxiliary variables for NHANES data. The areas of the bubbles are proportional to sample weights

4. Case Study: NHANES

In this section, we examine a regression of systolic blood pressure on the logarithm of blood lead level, age, and body mass index using a subset from the National Health and Nutrition Examination Survey (NHANES) 1999-2002. The subset used in this study has a sample size of 810, consisting of Mexican-American females aged 20 to 29. This sample does not have very skewed $\mathbf{Y}$ and $\mathbf{X}$ values, but involves clustering and stratification in the sampling design with a set of large and greatly varying sample weights. There are $n = 57$ PSUs nested in $H = 28$ strata, all but one of the strata having 2 PSUs. The average cluster size $\bar{m}$ is 14.21 persons. When applied to a clustered data set, the variance estimators in the survey-weighted diagnostic statistics need to take the design into account and the cutoffs

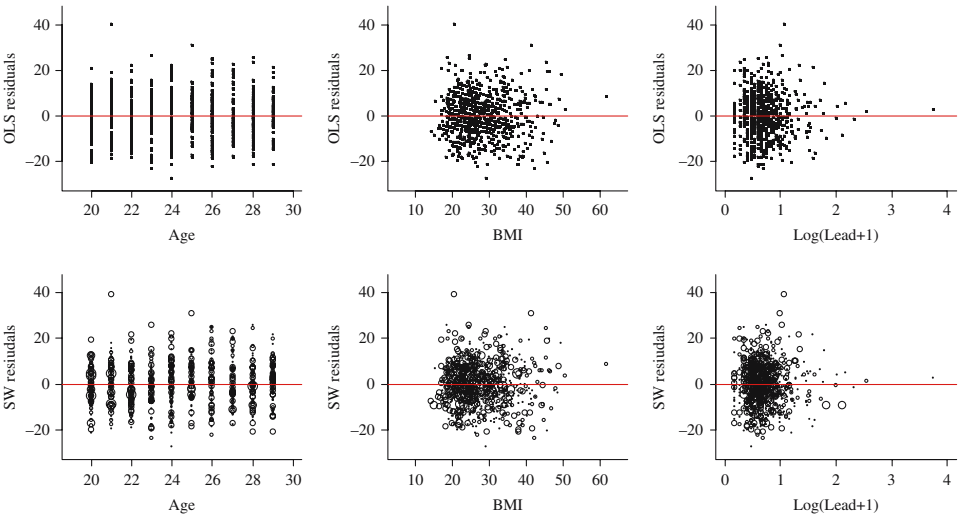

Fig. 2. OLS and SW residuals versus three auxiliary variables for NHANES data. Horizontal reference lines are drawn at zero

Table 2. OLS and SW parameter estimates from NHANES regression

	OLS Estimation			SW Estimation		
Independent Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	94.91***	3.11	30.55	99.79***	4.72	21.16
Age	0.02	0.11	0.14	−0.15	0.17	−0.87
BMI	0.45***	0.05	9.23	0.44***	0.07	5.88
Log(Lead+1)	1.03	0.99	1.04	0.89	1.28	0.70

*** Significant at level 0.001

for some of the statistics contain an estimate of ρ, which in Model (1) describes the correlation between the observations within the same cluster. The illustrative calculations in this study do not account for the fact that Mexican-American females are a domain within the full population whose sample size is random. This will tend to make SW variance estimates smaller than they would be if the domain feature was accounted for.

Table 1 gives the quantile values of the variables and sample weights used in the regression. Besides demonstrating the skewness and large range of sample weights, the table also shows that the distributions of BMI and the logarithm of the blood lead are skewed to the right. Since the minimum of the originally measured blood lead level is as small as 1, we added 1 to blood lead level before taking the logarithm to generate positive transformed values. (Adding 1 is often done to avoid taking the log of zero; this step was not strictly necessary here.) Note that using the untransformed value of blood lead would have resulted in more extreme **X** values. However, this type of modeling has previously been done using the log transformation (see Korn and Graubard 1999), and we follow that precedent here. Figures 1 and 2 respectively display plots of systolic blood pressure and residuals versus the three auxiliary variables. Table 2 reports the parameter estimates of the regressions with and without weights. The SW estimators produced slightly larger intercept and slightly smaller slope of BMI than the OLS ones. Both methods agree that

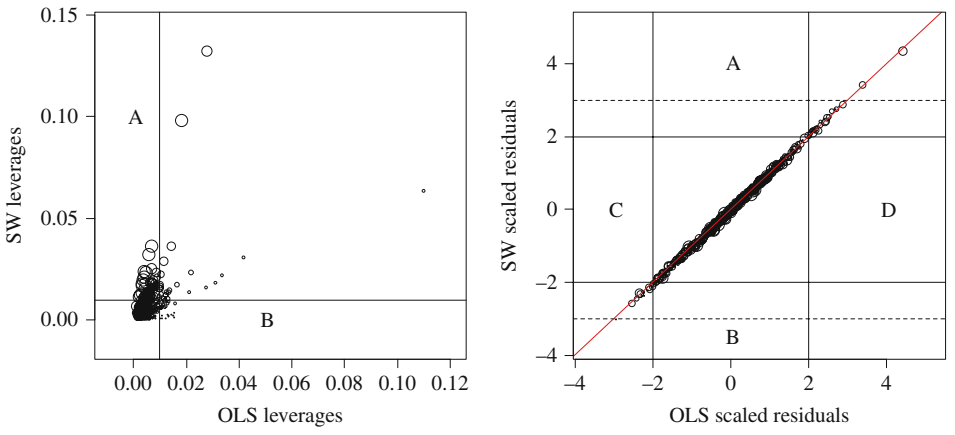

Fig. 3. Leverage and residual plots for NHANES data. In left-hand panel, A = points identified only by SW diagnostics; B = points identified only by OLS diagnostics; vertical and horizontal reference lines are drawn at $2p/n\tilde{m}$. In right-hand panel, A,B = points identified by SW but not OLS. C,D = points identified by OLS but not SW

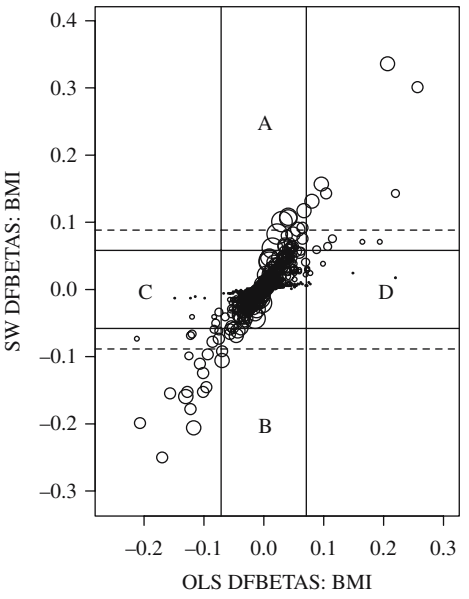

Fig. 4. DFBETAS Plot of BMI for NHANES Data. A,B = points identified by SW but not OLS. C,D = points identified by OLS but not SW

age and blood lead do not have significant effects in determining the systolic blood pressure. Therefore, in the following diagnostic analysis, we will only focus on the changes in the estimated coefficient of BMI.

For comparison, we applied both the OLS and the new SW diagnostic statistics, including leverages, residuals, DFBETAS, DFFITS, and modified Cook's distance, to the regression estimation. Since the sample weights were not separately provided at cluster level and at unit level, the parameters ρ and σ^2 in Model (1) were estimated using purely model-based estimators. Utilizing the `VARCOMP` procedure in SAS, we obtained $\hat{\rho} = 0.033$ and $\hat{\sigma}^2 = 82.09$. The design effect was estimated as $\sqrt{1 + \hat{\rho}(\bar{m} - 1)} = 1.2$. For the

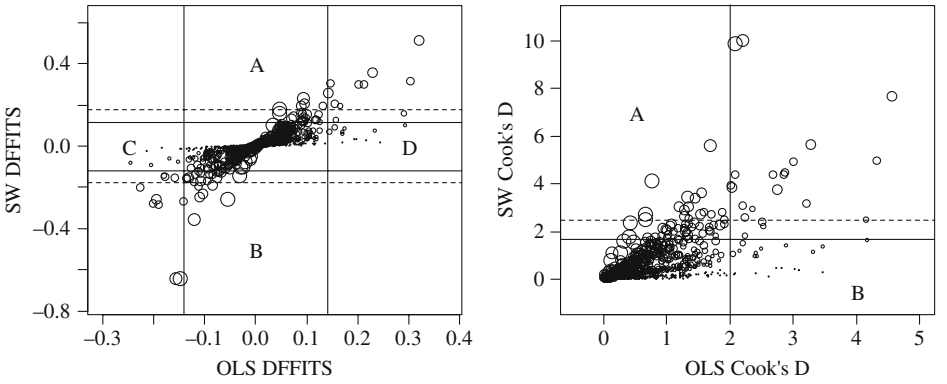

Fig. 5. DFFITS plot and modified Cook's distance plot for NHANES data. In left-hand panel A,B = points identified by SW but not OLS; C,D = points identified by OLS but not SW. In right-hand panel A = points identified by SW but not OLS; B = points identified by OLS but not SW

Table 3. Number of outliers identified and associated weight ranges for NHANES data

	Outliers identified by OLS only		Outliers identified by SW only	
Diagnostic statistics	Counts	Weight range	Counts	Weight range
Leverage	24	(875.5, 13,085.8)	85	(16,929.6, 103,831.2)
Residual	1	(2,730.1, 2,730.1)	8	(1,791.1, 36,955.3)
DFBETAS(BMI)	25	(1,773.5, 2,3677.5)	12	(32,451.1, 103,831.2)
DFFITS	21	(994.9, 17,366.9)	28	(2,9617.1, 103,831.2)
Modified Cook's D	21	(994.9, 17,366.9)	35	(21,194.0 103,831.2)

SW diagnostics, a strict criterion, 2, was used to construct cutoffs. For example, the cutoff of DFBETAS is $2\Big/\sqrt{n\bar{m}\left[1+\hat{\rho}(\bar{m}-1)\right]}$. The solid reference lines in the subsequent figures were drawn at the cutoff values of 2; dotted reference lines using the looser criterion of 3 are also drawn in the same graphs.

Figures 3 through 5 display the comparisons between the OLS and the SW diagnostic statistics. The range of the weights in the NHANES data set is extremely wide, with a minimum of 698.39 and a maximum of 103,831.17. Hence the SW diagnostics tend to identify more influential observations with large weights, whereas the OLS diagnostics tend to detect more points with small weights. The leverage plot (Figures 3), DFBETAS plot (Figure 4), and the modified Cook's distance plot (Figure 5) clearly show that the "identified by SW only" areas contain many big bubbles, but the "identified by OLS only" areas are filled with small dots. The residual plot is an exception in which the OLS and the SW residuals are very similar. This is mainly because none of the **Y** and **X** values in the data set are extremely outlying.

Table 3 numerically reports the weight discrepancies between the observations uniquely identified by either OLS or SW diagnostics. The leverage and modified Cook's distance are more sensitive to extreme sample weights compared to other diagnostic statistics. They tend to detect more influential points for survey data than the OLS approaches. Analysts may want to consider raising the cutoff values for these statistics in order not to overidentify influential points.

Table 4. Estimated slopes of BMI from full sample and reduced samples by different diagnostic approaches for NHANES data

	OLS estimation			SW estimation		
	BMI	SE	*t*	BMI	SE	*t*
Full sample	0.45***	0.05	9.23	0.44***	0.07	5.88
Leverages	0.39***	0.06	6.86	0.43***	0.08	5.23
Residuals	0.47***	0.04	10.50	0.47***	0.06	8.19
DFBETAS (BMI)	0.49***	0.05	9.51	0.46***	0.05	8.83
DFFITS	0.47***	0.05	9.76	0.45***	0.05	8.51
Modified Cook's D	0.47***	0.05	9.76	0.44***	0.05	8.74

*** Significant at level 0.001

The parameter estimates after outliers were removed are listed in Table 4. The difference between the OLS and SW estimates and the two diagnostic schemes is trivial. The removal of observations with large DFBETAS of BMI causes the largest change in the estimated slope of BMI. The SW estimates seem to be less affected by the removal of influential points than the OLS ones. Unlike the SMHO data analyzed in Li and Valliant (2011a), the NHANES data set does not contain many obviously extreme points, and outlying Y values can be large or small relative to other points. Hence the deletion of the identified outliers does not move the regression line dramatically.

5. Conclusion

By incorporating survey weights and design features, we constructed survey-weighted diagnostic statistics for clustered samples that are extensions of the conventional OLS diagnostics. Survey-weighted diagnostics may identify different points than OLS diagnostics as influential. An observation with moderate Y and $\mathbf{x}$ values may not be identified as influential by OLS approaches, but may be recognized as influential by SW methods if it is assigned an extreme sample weight. The diagnostics can serve as a guide to which points may be unusual. However, a diligent analyst should examine these points in detail to decide whether they are data entry errors, legitimate values that do not follow a core model, or can be explained in some other way, such as having extreme weights.

The techniques based on single-case deletion presented here may not function effectively when some outliers mask the effects of others. The modified forward search method (Atkinson and Riani 2000, 2004; Li and Valliant 2011b) is a partial solution to this problem since it can successfully identify an influential group of points whose members are not influential when examined singly.

A final caveat to the use of the diagnostics studied here is that some points may appear to be influential because the regression model itself is misspecified. Deleting them would be a mistake if the ability is lost to recognize that the model should be respecified, for example, as quadratic. Thus good practice will require using a combination of residuals and the other diagnostics studied here.

6. References

Atkinson, A.C., and M. Riani. 2000. *Robust Diagnostic Regression Analysis*. New York: Springer-Verlag.

Atkinson, A.C., and M. Riani. 2004. "The Forward Search and Data Visualization." *Computational Statistics* 19: 29–54.

Bates, D., M. Maechler, B. Bolker and S. Walker. 2014. "*lme4: Linear Mixed-Effects Models Using Eiqen and S4. R package version 1.1-7*." Available at: http://CRAN.R-project.org/package= lme4 (accessed February 2, 2015).

Belsley, D.A., R. E. Kuh, and R. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

Binder, D.A. 1983. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51: 279–292. DOI: http://dx.doi.org/10.2307/1402588

Chambers, R.L., A.H. Dorfman, and T.E. Wehrly. 1993. "Bias Robust Estimation in Finite Populations Using Nonparametric Calibration." *Journal of the American Statistical Association* 88: 268–277. DOI: http://dx.doi.org/10.1080/01621459.1993.10594319

Chambers, R.L. 1996. "Robust Case-Weighting for Multipurpose Establishment Surveys." *Journal of Official Statistics* 12: 3–32.

Chambers, R.L., and C.J. Skinner. 2003. *Analysis of Survey Data*. New York: John Wiley.

DuMouchel, W.H., and G.J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples." *Journal of the American Statistical Association* 78: 535–543. DOI: http://dx.doi.org/10.1080/01621459.1983.10478006

Elliott, M. 2007. "Bayesian Weight Trimming for Generalized Linear Regression Models." *Survey Methodology* 33: 23–34.

Fuller, W.A. 2002. "Regression Estimation for Survey Samples." *Survey Methodology* 28: 5–23.

Graubard, B.I., and E.L. Korn. 1996. "Modelling the Sampling Design in the Analysis of Health Surveys." *Statistical Methods in Medical Research* 5: 263–281. DOI: http://dx.doi.org/10.1177/096228029600500304

Henry, K.A., and R. Valliant. 2012. "Methods for Adjusting Survey Weights When Estimating a Total." In *Proceedings of the Federal Committee on Statistical Methodology*, January 10–12. Washington, DC. Available at: http://fcsm.sites.usa.gov/files/2014/05/Henry_2012FCSM_V-A.pdf (accessed February 2, 2015)

Korn, E.L., and B.I. Graubard. 1999. *Analysis of Health Surveys*. New York: Wiley.

Korn, E.L., and B.I. Graubard. 2003. "Estimating Variance Components by Using Survey Data." *Journal of Royal Statistical Society B* 65: 175–190. Part 1. DOI: http://dx.doi.org/10.1111/1467-9868.00379

Kott, P.S. 1991. "A Model-Based Look at Linear Regression with Survey Data." *American Statistician* 45: 107–112. DOI: http://dx.doi.org/10.1080/00031305.1991.10475779.

Li, J., and R. Valliant. 2009. "Survey Weighted Hat Matrix and Leverages." *Survey Methodology* 35: 15–24.

Li, J., and R. Valliant. 2011a. "Linear Regression Influence Diagnostics for Unclustered Survey Data." *Journal of Official Statistics* 27: 99–119.

Li, J., and R. Valliant. 2011b. "Detecting Groups of Influential Observations in Linear Regression using Survey Data—Adapting the Forward Search Method." *Pakistan Journal of Statistics* 27: 507–528.

Liao, D., and R. Valliant. 2012a. "Variance Inflation Factors in the Analysis of Complex Survey Data." *Survey Methodology* 38: 53–62.

Liao, D., and R. Valliant. 2012b. "Condition Indexes and Variance Decompositions for Diagnosing Collinearity in Linear Model Analysis of Survey Data." *Survey Methodology* 38: 189–202.

Longford, N.T. 1995. *Models for Uncertainty in Educational Testing*. New York: Springer-Verlag.

Miller, R.G., Jr. 1974. "An Unbalanced Jackknife." *The Annals of Statistics* 2: 880–891.

Pfeffermann, D., and D.J. Holmes. 1985. "Robustness Considerations in the Choice of Method of Inference for the Regression Analysis of Survey Data." *Journal of the Royal Statistical Society A* 148: 268–278. DOI: http://dx.doi.org/10.2307/2981971

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society B* 60: 23–40. DOI: http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00106/abstract

Potter, F.A. 1988. "Survey of Procedures to Control Extreme Sampling Weights." In *Proceedings of the Section on Survey Research Methods: American Statistical Association*, 453–458. Available at: http://www.amstat.org/sections/SRMS/proceedings/.

Potter, F.A. 1990. "Study of Procedures to Identify and Trim Extreme Sample Weights." In *Proceedings of the Survey Research Methods Section: American Statistical Association*, 225–230. Available at: http://www/amstat.org/sections/SMRM/proceedings/.

Pukelsheim, F. 1994. "The Three Sigma Rule." *The American Statistician* 48: 88–91.

Scott, A.J., and D. Holt. 1982. "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods." *Journal of the American Statistical Association* 77: 848–854.

Skinner, C.J., D. Holt, and T.M.F. Smith (eds.). 1989. *Analysis of Complex Surveys*. New York: Wiley.

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

Wolter, K. 2007. *Introduction to Variance Estimation*. New York: Springer.

Zaslavsky, A., N. Schenker, and T. Belin. 2001. "Downweighting Influential Clusters in Surveys: Application to the 1990 Post Enumeration Survey." *Journal of the American Statistical Association* 96: 858–869.

Received August 2013
Revised May 2014
Accepted September 2014