Journal of Official Statistics, Vol. 27, No. 1, 2011, pp. 99–119

Linear Regression Influence Diagnostics for Unclustered Survey Data

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

Diagnostics for linear regression models have largely been developed to handle nonsurvey data. The models and the sampling plans used for finite populations often entail stratification, clustering, and survey weights. In this article we adapt some influence diagnostics that have been formulated for ordinary or weighted least squares for use with unclustered 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 in an empirical study where the values of weights, response variables, and covariates vary substantially.

Key words: Complex sample; Cook's D; DFBETAS; DFFITS; influence; outlier; residual analysis.

1. Introduction

Diagnostics for identifying influential points are staples of standard regression texts like Cook and Weisberg (1982), Neter et al. (1996) and Weisberg (2005). These diagnostics have been developed for linear regression models fitted with nonsurvey data. The diagnostic tools provided by current, popular software packages are generally based on ordinary or weighted least squares (OLS or WLS) regression and do not account for stratification, clustering, and survey weights that are features of data sets collected in complex sample surveys. The OLS/WLS diagnostics can mislead users either because survey weights are ignored, or the variances of model parameter estimates are estimated incorrectly by the standard procedures. Hence, the goal of this article is to adapt and extend some of the standard regression diagnostics to the survey setting, and, where necessary, develop new ones.

A fundamental question is whether a few unusual points will be troublesome in a survey data set where the number of observations may be in the hundreds or thousands. In fact, examples from real surveys show that there is a need for influence diagnostics since a small number of the sampled units with extreme values can play a crucial role in estimators and their variances. For instance, in 1986, the Joint Economic Committee of the

[1] Westat, 1650 Research Boulevard, Rockville, MD 20850, U.S.A. Email: JaneLi@westat.com
[2] Joint Program in Survey Methodology, University of Maryland and University of Michigan, 1218 Lefrak Hall, College Park MD 20742, U.S.A. Email: rvalliant@survey.umd.edu

Acknowledgments: This material is based, in part, upon work supported by the U.S. National Science Foundation under Grant No. 0617081. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

U.S. Congress released a study indicating a sharp increase in the percentage of wealth held by the most affluent families in America. The richest 0.5% of families was estimated to hold 35% of the wealth in 1983, whereas in 1963 this proportion was 25%. The finding was proved to be wrong because a respondent with a very large weight was recorded to have $200 million in wealth attributed to him when the correct number was $2 million (Ericksen 1988). The estimated share of wealth held by the richest 0.5% of families dropped to 27% after the figure was corrected.

As in other statistical disciplines, outliers have been a well-known problem in survey sampling (e.g., see Lee 1995). Usually outliers feature extreme values that may be substantially different from the bulk of the data. Chambers (1986) characterized outliers into two basic types: nonrepresentative and representative. The former means the value for a sample unit is incorrect or the value is unique to a particular population unit, whereas the latter refers to cases in which the values are correct and there are others like them in the nonsample part of the population. There are diverse reasons for survey data containing influential observations, such as editing error, observation error, or simply a skewed variable. As Smith (1987) pointed out, "individual values can be influential in randomization inference either when they are included in the sample or when they are not in the sample," and "diagnostics are useful in the former case." A few nonsample, nonrepresentative outliers, for example, can have a large effect on the error of an estimated total but cannot be identified by diagnostics. Our analyses focus on discovering the sample points that affect estimators of linear regression model parameters.

Another feature of survey data is that extreme values of response variables or covariates and influential values may not necessarily refer to the same observations due to sizes of sample weights. The distinction between the two concepts has been noted by some survey researchers (see Gambino 1987; Srinath 1987).

The premise in this research is that an analyst will be looking for a linear regression model that fits reasonably well for the bulk of the finite population. We have in mind two general goals. First, the influence diagnostics should allow the analyst to identify points that may not follow that model and have an influence on the size of estimated model parameters, their estimated standard errors, or both due to outlying values of dependent or independent variables. Second, the diagnostics should identify points that are influential because of the size of the survey weights. Weights may be extreme because of the way the sample was selected or because of subsequent nonresponse or calibration adjustments.

Conventional model-based influence diagnostics mainly use the technique of row deletion, determining if the fitted regression function is dramatically changed when one or multiple observations are discarded. The statistics which are widely adopted include DFBETAS, DFFITS, and Cook's Distance, among others (e.g., see Neter et al. 1996). We review some of these in Section 2.

In developing diagnostics for survey data, a basic question is which theory should be used for motivation – design-based or model-based. The standard statistics listed above do not have immediate application to randomization inference for sample surveys. As Brewer and Särndal (1983) noted, the idea of robustness to departures from an assumed model does not fit naturally into a purely design-based framework, because models are not used directly in inference. However, the consideration of a model is needed to motivate the use of diagnostic statistics in finite population inference. The goal of inference will be to develop procedures that permit "good" estimates of parameters for the core model, that is, one that fits for most of a finite population. A byproduct of this thinking is that nonsample points that do not follow the core model are not a concern in contrast to the problem of estimating finite population totals where such representative outliers must be considered. By omitting influential points, ideally, less biased and more stable estimates of underlying model parameters will result.

A key point to bear in mind is that the measures that are in the literature for nonsurvey regressions and the ones we present mainly have heuristic justifications only. There is limited distribution theory to support the setting of cutoff values for statistics that gauge whether a point is influential or not. Nonetheless, the measures in this article give some practical, exploratory tools for identifying points for further examination. After reviewing some standard diagnostics in Section 2, we sketch in Section 3 the method of pseudo-maximum likelihood, which is used for estimating linear model parameters from survey data. Section 4 adapts several diagnostics for use with complex survey data. An empirical study in Section 5 compares and contrasts the information conveyed by OLS and survey-weighted diagnostics. Section 6 is a conclusion.

2. Review of Traditional Techniques

The conventional diagnostic techniques are applied in the context of the regression model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i \sim (0, \sigma^2 v_i);\ i = 1, \ldots, n \tag{1}$$

where Y_i is a response variable for unit i, $\mathbf{x}_i$ is a p-vector of fixed covariates, $\boldsymbol{\beta}$ is a fixed but unknown parameter, the ε_i's are independent random variables with mean 0 and variance $\sigma^2 v_i$. Of course, there are more elaborate random and mixed effect models that may be appropriate in some applications, but we will not consider those here. In matrix-vector notation, Model (1) is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (0, \sigma^2 \mathbf{V})$ with $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, and $\mathbf{V} = diag(v_i)$ is an $n \times n$ diagonal matrix. The WLS estimator of $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$. When $\mathbf{V} = \mathbf{I}$, the $n \times n$ identity matrix, this reduces to the OLS estimator, $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. We review some OLS diagnostics in the remainder of Section 2.

2.1. *Leverages and Residuals*

In conventional model diagnostics, the residuals, $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$, and the hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, are the measures used to identify outlying $\mathbf{Y}$ and $\mathbf{X}$ values, respectively. The diagonal element $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ of the hat matrix is called the leverage of the ith case. A leverage value is usually considered as large if it is more than twice its mean, p/n. The residuals, $e_i = Y_i - \mathbf{x}_i^T\mathbf{b}$, are often rescaled relative to their standard errors. The ratio of e_i to $s_{e_i} = \sqrt{s^2(1 - h_{ii})}$, where $s^2 = \sum_{i=1}^n e_i^2/(n - p)$ is the mean squared error (MSE), is called the internally studentized residual and denoted by r_i. Replacing s^2 with $s^2(i)$, the mean squared error when the ith case is omitted in fitting the regression function, we obtain an externally studentized residual

$$r_i^* = \frac{e_i}{s(i)\sqrt{1 - h_{ii}}}$$

which follows the t distribution with $n - p - 1$ degrees of freedom assuming that Model (1) holds with $v_i = 1$ and with the additional assumption that the errors are normal.

2.2. Influence on Regression Coefficients: DFBETA and DFBETAS

DFBETA, the change in parameter estimates after deleting the ith observation, can be written as $DFBETA \equiv \mathbf{b} - \mathbf{b}(i) = \mathbf{A}^{-1}\mathbf{x}_i e_i/(1 - h_{ii})$, where $\mathbf{A} = \mathbf{X}^T\mathbf{X}$. If we let $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = (c_{ji})_{p\times n}$, then the jth element of the DFBETA vector is $b_j - b_j(i) = c_{ji}e_i/1 - h_{ii}$ (Belsley, Kuh, and Welsch 1980; referred to as BKW subsequently). If the $\mathbf{X}$'s are uniformly bounded, then $c_{ji} = O(n^{-1})$. BKW suggest that the changes in the estimated regression coefficients are often most usefully assessed relative to the model variance of $\mathbf{b}$. A scaled measure of the change can be defined as the following:

$$DFBETAS_{ij} = \frac{b_j - b_j(i)}{s(i)\sqrt{(\mathbf{X}^T\mathbf{X})_{jj}^{-1}}} = \frac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c_{jk}^2}} \frac{r_i^*}{\sqrt{1 - h_{ii}}}$$

where $(\mathbf{X}^T\mathbf{X})_{jj}^{-1}$ is the (jj)th element of $(\mathbf{X}^T\mathbf{X})^{-1}$. The denominator of DFBETAS$_{ij}$ is analogous to the estimated standard error of $\mathbf{b}$ with the sample standard error s replaced by the delete-one version $s(i)$. The DFBETAS statistic is the product of a quantity of order $n^{-1/2}$, a t distributed random variable (the externally standardized residual r_i^*), and a quantity, $(1 - h_{ii})^{-1/2}$, that approaches 1 (assuming $h_{ii} \rightarrow 0$). BKW propose a cutoff point of $2/\sqrt{n}$ to identify influential cases. Thus, if all the observations in the sample follow an underlying normal model, the $\mathbf{X}$'s are bounded, and the leverages are small, roughly 95% of the observations will have a DFBETAS statistic less than $2/\sqrt{n}$ in absolute value. In some samples, especially small or moderate size ones, this statement is less accurate since h_{ii} may not be negligible and the term involving c_{ji} may not be near $n^{-1/2}$. DFBETAS is somewhat cumbersome to work with because an analyst must examine pn values. For each observation i, there are p DFBETAS – one for each parameter.

2.3. Influence on Fitted Values: DFFIT and DFFITS

DFFIT is a statistic that summarizes the change in predicted values when an observation is deleted, with the advantage that it does not depend on the particular coordinate system used to form the regression model. DFFITS is constructed by rescaling DFFIT by the estimated standard deviation of the predicted value, with the sample standard error s replaced by the delete-one version $s(i)$. DFFITS can be expressed as the product of a t distributed random variable and a function of the leverage:

$$DFFITS_i \equiv \frac{\hat{Y}_i - \hat{Y}_i(i)}{s(i)\sqrt{h_{ii}}} = \frac{\mathbf{x}_i^T(\mathbf{b} - \mathbf{b}(i))}{s(i)\sqrt{h_{ii}}} = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} r_i^*$$

A large value of DFFITS indicates that the observation is very influential in its neighborhood of the $\mathbf{X}$ space. A general cutoff to consider is 2; a size-adjusted cutoff recommended by BKW is $2\sqrt{p/n}$, where p/n is the mean leverage.

2.4. Cook's Distance

Cook's distance provides an overall measure of the combined impact of an observation on all of the estimated regression coefficients **b** (e.g., see Cook 1977 and Weisberg 2005). It is motivated by considering the confidence region of $\boldsymbol{\beta}$, which at level $100(1-\alpha)\%$ is given by those values $\mathbf{b}^*$ satisfying

$$(\mathbf{b}^* - \mathbf{b})^T \mathbf{X}^T \mathbf{X}(\mathbf{b}^* - \mathbf{b})/ps^2 \leq F(1-\alpha; p, n-p)$$

Using the same structure, Cook's distance measure D_i was proposed as

$$D_i = (\mathbf{b}(i) - \mathbf{b})^T \mathbf{X}^T \mathbf{X}(\mathbf{b}(i) - \mathbf{b})/ps^2$$

This is a measure of the distance from $\mathbf{b}(i)$ to $\mathbf{b}$. If $\mathbf{b}(i)$ and $\mathbf{b}$ are relatively far from each other, this means that unit i has a substantial effect on the full sample estimate. Large values of D_i indicate observations that are influential on joint inferences about all the parameters in the linear model. Although there is no formal distributional theory for D_i, the standard procedure, originally suggested by Cook, is to compare D_i to the percentile values of the $F(1-\alpha; p, n-p)$ distribution to make a judgment on influence.

Cook's D can also be written in terms of the residual and leverage for unit i:

$$D_i = \frac{e_i^2 h_{ii}}{ps^2(1-h_{ii})^2} = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$$

Atkinson (1982) suggested replacing s^2 by the deletion estimate $s^2(i)$, scaling the statistic by the average leverage *p*/*n*, and then taking the square root to give a residual like quantity. The resulting modified Cook statistic is

$$D_i^* = \left(\frac{n-p}{p}\right)^{1/2}\left(\frac{h_{ii}}{(1-h_{ii})^2}\frac{e_i^2}{s^2(i)}\right)^{1/2} = \left(\frac{n-p}{p}\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}\left|r_i^*\right|$$

where r_i^* is the externally studentized residual defined earlier. If n is large, a heuristic cutoff for the modified Cook's distance is 2 because r_i^* has a t distribution.

3. Linear Regression Estimation With Complex Survey Data

One method of estimating parameters in linear regression using complex survey data is the pseudo maximum likelihood (PML) approach, outlined by Skinner et al. (1989), following ideas of Binder (1983). The first step of this approach is to write down and maximize the likelihood when all finite population units are observed. Suppose that the underlying structural model is the fixed-effects linear model given by (1). The pseudo maximum likelihood estimator (PMLE) of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{Y}$.

Model (1) may be appropriate for a population in which single-stage sampling is used, e.g., hospitals or establishments from a business frame, or persons from an organizational membership list. In populations where there is natural clustering, like students within schools or households in neighborhoods, a model that accounts for intracluster correlation among different units may be more realistic.

The regression estimator $\hat{\boldsymbol{\beta}}$, which incorporates the sample weights $\mathbf{W}$, is approximately design unbiased for the finite population parameter $\mathbf{B} = (\mathbf{X}_N^T\mathbf{V}_N^{-1}\mathbf{X}_N)^{-1}\mathbf{X}_N^T\mathbf{V}_N^{-1}\mathbf{Y}_N$, where

$\mathbf{Y}_N = (Y_1, \ldots, Y_N)^T$, $\mathbf{V}_N = diag(v_1, \ldots, v_N)$, and $\mathbf{X}_N^T = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ assuming that the weights (w_i's) are constructed to produce design-unbiased estimates of finite population totals. This estimator is also unbiased for the superpopulation slope $\boldsymbol{\beta}$ in model (1), regardless of whether $\mathbf{V}$ is specified correctly or not. When the population is large, the finite population parameter $\mathbf{B}$ should be close to the model parameter $\boldsymbol{\beta}$ if the model is correctly specified, and therefore a design-based estimator of $\mathbf{B}$ should also estimate $\boldsymbol{\beta}$. If we assume $\mathbf{V} = \mathbf{I}$, the PMLE reduces to $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, which is the case we study in the remainder of this article. This estimator will be referred to as a survey weighted (SW) estimator in the following discussion and is the one usually computed by software packages that handle survey data.

Researchers who advocate model-based approaches may argue that the sample design should have no effect in regression estimation as long as the design is ignorable and the observations in the population and sample really follow the model. In that case, an OLS estimator or WLS estimator that uses only $\mathbf{V}^{-1}$ (not $\mathbf{W}$) can be used to infer about the model parameters. However, with survey data a theoretically derived model rarely holds for all observations. First, the model may not be appropriate for every subgroup in the population; second, some relevant explanatory variables may not be measured in the survey; third, the true relations among the variables may not be exactly linear. In addition, informative nonresponse can distort the model relationship observed in the sample compared to that in the population because of the dependence of response on variables of interest.

Using sampling weights in a regression can provide a limited type of robustness to model misspecification. From a model-based perspective, Rubin (1985), Smith (1988), and Little (1991) argue that the sampling weights are useful as summaries of covariates which describe the sampling mechanism. Pfeffermann and Holmes (1985), DuMouchel and Duncan (1983), and Kott (1991) claim that the estimators using sampling weights are less likely to be affected if some independent variables are not included in the model. Thus, even if one is doing model-based analysis, using weights may provide some protection against bias due to omitting certain regressors.

Another advantage of using the weighted estimators is the ability to say we are estimating a population quantity with the price of generally larger estimated variances than for OLS. If the working model is good, we expect that the point estimators $\hat{\boldsymbol{\beta}}$ and $\mathbf{b}$ should be similar. However, if the model is misspecified, survey-weighted and OLS estimates can be far apart, as illustrated in Korn and Graubard (1995). Use of the weights also seems, at first glance, to be consistent with our goal of finding a model that fits for most of the population. With that reasoning, a unit with a large weight would represent a large portion of the population and using the weight would reflect that. Although this intuitive explanation of the meaning of a weight is sometimes reasonable, this is not always true. A large weight may be the result of an extreme nonresponse or calibration adjustment that reduces neither bias nor variance. It might also be a mistake, as in the example of the misplaced decimal point in the example from Ericksen (1988).

Here we address cases where analysts will use survey weights to estimate regression models. The diagnostics to be developed account for the effects of these weights with some awareness of the possibility that large weights can be a source of influence. The diagnostics should allow such cases to be identified so that an analyst can decide whether to include them or not.

4. Adaptations of Standard Techniques to Survey Regressions

Although survey weights are used in PMLE's, implying that an analyst may be interested in design-based properties, explicitly appealing to models is necessary to motivate diagnostics. In this section, we examine residuals and extensions of DFBETAS, DFFITS, and Cook's D to survey data, relying on models to justify the forms of the diagnostics and cutoffs for identifying influential points.

4.1. Variance Estimators

To construct several of the diagnostics, an estimator of the variance of the SW regression parameter estimator is required. We use $v(\hat{\boldsymbol{\beta}})$ and $v(\hat{\beta}_j)$ to denote a general estimator that is appropriate from a design-based or model-based point of view. To calculate the diagnostics, any of several options can be used for $v(\hat{\boldsymbol{\beta}})$ and $v(\hat{\beta}_j)$. When first-stage units are selected with replacement, the sandwich estimator (Binder 1983) or replication estimators like the jackknife can be constructed that are consistent and approximately design-unbiased for single-stage or multistage sampling. These estimators are also consistent and approximately model-unbiased under model (1) (e.g., see Li 2007). There are also purely model-based estimators of the model variance of the PMLE $\hat{\boldsymbol{\beta}}$. For example, an estimator of the model-variance under Model (1) is

$$v_M(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \mathbf{A}^{-1} \left(\sum_{i=1}^{n} w_i^2 \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{A}^{-1} = \hat{\sigma}^2 \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W}^2 \mathbf{X} \mathbf{A}^{-1} \tag{2}$$

where $\mathbf{A} = \mathbf{X}^T \mathbf{W} \mathbf{X}$ and

$$\hat{\sigma}^2 = \sum_{i \in s} w_i e_i^2 / (\hat{N} - p) \tag{3}$$

with $\hat{N} = \sum_{i \in s} w_i$ and $e_i = Y_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$. In the preceding formulas, s denotes the set of sample units, rather than a sample variance as in Section 2. The estimator $\hat{\sigma}^2$ is approximately design-unbiased for $\sum_{i=1}^{N} e_{N_i}^2 / N$ with $e_{N_i} = Y_i - \mathbf{x}_i^T \mathbf{B}$ when $p \ll N$. The estimator $\hat{\sigma}^2$ is also approximately model-unbiased for σ^2 and reduces to the usual OLS estimator when $w_i \equiv 1$.

The model-based estimator above is useful because it explicitly shows the estimates of model parameters. With some simplifications, described later, (2) is helpful in setting heuristic cutoffs that can be used with the diagnostics. There are variance estimators that are more robust to departures from Model (1) (e.g., see Valliant et al. 2000) but are not convenient for setting cutoffs.

4.2. Leverages

The hat matrix associated with the PMLE $\hat{\boldsymbol{\beta}}$ is $\mathbf{H} = \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W}$. The leverages are the diagonal of the hat matrix and are equal to $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i w_i$. Li and Valliant (2009) cover their properties in detail; we sketch them here since leverages will be part of the empirical study in Section 5. Leverages depend on covariates and weights but are not affected by variation in Y. A leverage can be large, and, as a result, influential on predictions, when an $\mathbf{x}_i$ is considerably different from the weighted average, $\bar{\mathbf{x}}_w = \sum_{i \in s} w_i \mathbf{x}_i / \sum_{i \in s} w_i$, or when the weight w_i is much different from their sample average, $\bar{w} = \sum_s w_i / n$.

4.3. Residual Analysis

Standardizing residuals is helpful so that their variance is approximately 1. In the OLS case, a residual is scaled either by $\sqrt{\text{MSE}}$ or by its estimated standard error. Under Model (1), the residual for unit i based on the PMLE is $e_i = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ and its model variance is $E_M\left(e_i^2\right) = \sigma^2\left[(1-h_{ii})^2 + \sum_{i' \neq i} h_{ii'}^2\right]$. Since $h_{ii'} = O(n^{-1})$, (e.g., see Li and Valliant 2009), the term in the brackets has the form $1 + o(1)$, and $E_M\left(e_i^2\right) \doteq \sigma^2$. We can standardize the residual for unit i as $e_i/\hat{\sigma}$ and compare it to percentiles from the distribution of a standard normal random variable. An ad hoc alternative would be to use a t-distribution with $n-p$ degrees of freedom as the reference distribution for small or moderate size samples. If e_i is not normal, the Gauss inequality (Pukelsheim 1994) is useful for setting a cutoff value:

> *Gauss inequality*: If a distribution 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 $\tau^2 = E[(X-\mu_0)^2]$.

According to Model (1) the residual has a symmetric distribution with its mode and mean at zero. Setting $\tau^2 = \sigma^2 = \text{var}(X)$, the Gauss inequality with $r = 2\sigma$ implies that the absolute value of a residual has about 90% probability to be less than twice its standard deviation and with $r = 3\sigma$ about 95% probability to be less than three times its standard deviation. If we rescale the residuals by a consistent estimate of σ, we can use either 2 as a loose cutoff or 3 as a strict one to identify outlying residuals. These cutoffs are arbitrary and are only intended as a way of deciding which points should be examined more closely. Depending on an analyst's preference, larger cutoffs could be used which would result in fewer points being scrutinized.

Appealing to a model is necessary when analyzing residuals because it is not feasible to define the distribution of residuals from the design-based point of view, even asymptotically. For example, in single-stage sampling, $e_i = Y_i(1-h_{ii}) + \sum\limits_{i' \neq i \in s} h_{ii'}Y_{i'}$. Although the second term, $\sum\limits_{i' \neq i \in s} h_{ii'}Y_{i'}$, is a linear combination of the $Y_{i'}$'s, the first, which is specific to unit i, is not. Therefore, a large sample central limit result for repeated sampling does not apply to e_i, the residual for a specific unit. However, if we approach the analysis with a working model in mind, plots of residuals are helpful in highlighting data points suspected of unduly affecting the fit of regression. For instance, plots of observed Y's or residuals against predicted values are still useful.

4.4. DFBETAS

Taking the sampling weights $\mathbf{W}$ into consideration, $DFBETA_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = \mathbf{A}^{-1}\mathbf{x}_i e_i w_i/(1-h_{ii})$. This equality uses the same calculation needed for the jackknife variance estimator (see e.g., Valliant et al. 2000, Section 5.4.2). Although the formula for the DFBETA statistic looks very much like the one in the OLS case, there are differences in both numerator and denominator because sample weights are involved in the leverages and residuals. However, the formulas have exactly the same form as the one for WLS with weights inversely proportional to model variances. To create a complex sample version of DFBETAS, we need to divide DFBETA by an estimate of the standard error of $\hat{\boldsymbol{\beta}}$ that accounts for unequal weighting, stratification, and other design complexities.

Using $DFBETA_{ij} = (\mathbf{A}^{-1}\mathbf{x}_i e_i w_i)_j/(1 - h_{ii}) = c_{ji}e_i/(1 - h_{ii})$ and a variance estimator, $v(\hat{\beta}_j)$, for $\hat{\beta}_j$, a scaled statistic DFBETAS can be constructed as in the OLS case. We propose a specification of DFBETAS statistic as

$$DFBETAS_{ij} = \frac{c_{ji}e_i/(1 - h_{ii})}{\sqrt{v(\hat{\beta}_j)}}$$

The purely model-based estimator in (2), $v_M(\hat{\beta}_j) = \hat{\sigma}^2[\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}\mathbf{A}^{-1}]_{jj} = \hat{\sigma}^2\sum_{i=1}^{n} c_{ji}^2$ where c_{ji}^2 is the jjth element of $\mathbf{A}^{-1}w_i^2\mathbf{x}_i\mathbf{x}_i^T\mathbf{A}^{-1}$, is convenient for motivating cutoff values:

$$DFBETAS_{ij} = \frac{c_{ji}}{\sqrt{\sum_{i'=1}^{n} c_{ji'}^2}} \cdot \frac{e_i}{\hat{\sigma}} \cdot \frac{1}{1 - h_{ii}} \tag{4}$$

Using the order conditions $c_{jk} = O(n^{-1})$ and $h_{ii} = O(n^{-1})$, we approximate the DFBETAS statistic as the product of two terms, $DFBETAS_{ij} \doteq O(n^{-1/2}) \cdot N(0, 1)$. The first term, with an order of $n^{-1/2}$, can be approximated by $n^{-1/2}$ when the sampled units have similar $\mathbf{X}$ values and weights. An observation i may be identified as influential on the estimation of $\hat{\beta}_j$ if $|DFBETAS_{ij}| \geq z/\sqrt{n}$ for $z = 2$ or 3. An ad hoc alternative would be to use a cutoff of $t_{0.025}(n - p)/\sqrt{n}$ where $t_{0.025}(n - p)$ is the 97.5 percentile of the t-distribution with $n - p$ degrees of freedom.

4.5. DFFITS

Multiplying the DFBETA statistic by the $\mathbf{x}_i^T$ vector, we obtain the measure of change in the ith fitted value due to the deletion of the ith observation, $DFFIT_i = \hat{Y}_i - \hat{Y}_i(i) = \mathbf{x}_i^T(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)) = h_{ii}e_i/(1 - h_{ii})$. In general, the scaled version is defined as

$$DFFITS_i = \frac{h_{ii}e_i/(1 - h_{ii})}{\sqrt{v(\hat{\beta}_j)}}$$

where $v(\hat{\beta}_j)$ is appropriate to the design and/or model. The model variance is again convenient for motivating cutoffs.

The model variance of $\hat{Y}_i$ is $V_M(\hat{Y}_i) = \sigma^2(\mathbf{H}\mathbf{H}^T)_{ii} = \sigma^2\sum_{i'} h_{ii'}^2$, which is estimated by $v_M(\hat{Y}_i) = \hat{\sigma}^2\sum_{i'} h_{ii'}^2$. In OLS, $\sum_{i'} h_{ii'}^2 = h_{ii}$ because $\mathbf{H}\mathbf{H}^T = \mathbf{H}$ when $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, but this simplification does not occur when $\mathbf{H}$ contains the survey weights. Under single-stage sampling and Model (1), $DFFIT_i$ is divided by the square root of $v_M(\hat{Y}_i)$ and rearranged as follows:

$$DFFITS_i = \frac{h_{ii}}{1 - h_{ii}} \frac{e_i}{\hat{\sigma}} \frac{1}{\sqrt{\sum_{i'=1}^{n} h_{ii'}^2}}$$

When the sample weights do not have a large variation, we can use the rough approximation $\sum_{i'} h_{ii'}^2 \approx h_{ii}$. Because the mean of the leverages is p/n (see Valliant et al. 2000, Lemma 5.3.1), we set the cutoff value to be $z\sqrt{p/n}$ ($z = 2$ or 3) for using DFFITS to determine the influential observations.

4.6. Distance Measure (Extended and Modified Cook's Distance)

A measure of distance from $\hat{\boldsymbol{\beta}}(i)$ to $\hat{\boldsymbol{\beta}}$ for survey data can be constructed similar to a Wald Statistic, depending on the regression model of interest and the sampling design for the survey data. We propose a statistic based on the standard Cook's D and name it the extended Cook's Distance in our study. The statistic is

$$ED_i = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))^T[v(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)) \tag{5}$$

If $\hat{\boldsymbol{\beta}}(i)$ were replaced by an arbitrary value $\mathbf{b}_0$, (5) would have the form of a confidence ellipsoid. The new statistic ED_i can be compared to a chi-square distribution. If ED_i were exactly equal to the $(1-\alpha) \times 100\%$ quantile of the chi-square distribution with p degrees of freedom, then the deletion of the ith case would move the estimate of $\boldsymbol{\beta}$ to the edge of a $100(1-\alpha)\%$ confidence ellipsoid based on the complete data. A large value of this quadratic form indicates that the ith observation is likely to be influential in determining joint inferences about all the parameters in the regression model. Another formulation of the extended Cook's Distance can be derived from the Wald F statistic (Korn and Graubard 1990) as

$$ED_i' = \frac{n-p+1}{np}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))^T[v(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))$$

and its value can be compared with quantiles from an F distribution. How well the distribution of ED_i' can be approximated by an F depends on the nearness of the distribution of the residuals to normality.

Like the Cook's Distance, the proposed extended Cook's Distance statistic is related to the sample size in order of magnitude. If the weights are $O(N/n)$, then $\mathbf{A} = O(N)$ elementwise, and $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = \mathbf{A}^{-1}\mathbf{x}_i e_i w_i/(1-h_{ii}) = O(n^{-1})$. Thus, $ED_i = O(n^{-1})$ and, when n is large, very few observations can be identified to be influential even if small tail percentiles of F and chi-square statistics are adopted as cutoffs. Following Atkinson (1982), we modify the proposed extended Cook's Distance to solve this problem. Suppose the working model is (1). Then, using the model-based variance estimator in (2),

$$\begin{aligned} ED_i &= (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))^T\left[v_M(\hat{\boldsymbol{\beta}})\right]^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)) \\ &= \left(\frac{e}{\hat{\sigma}}\right)^2 \frac{1}{(1-h_{ii})^2} w_i \mathbf{x}_i^T \mathbf{A}^{-1}\left[\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}\mathbf{A}^{-1}\right]^{-1}\mathbf{A}^{-1}\mathbf{x}_i w_i \\ &= \left(\frac{e}{\hat{\sigma}}\right)^2 \frac{1}{(1-h_{ii})^2} w_i \mathbf{x}_i^T \left[\mathbf{X}^T\mathbf{W}^2\mathbf{X}\right]^{-1}\mathbf{x}_i w_i \end{aligned} \tag{6}$$

Under some reasonable conditions

$$w_i \mathbf{x}_i^T[\mathbf{X}^T\mathbf{W}^2\mathbf{X}]^{-1}\mathbf{x}_i w_i = O(n^{-1})$$

and, if the weights are not too variable, this quantity has a mean of approximately p/n. Hence, we suggest that an analyst take the square root of the extended Cook's D statistic and rescale the root by $(p/n)^{-1/2}$. The statistic $MD_i = \sqrt{nED_i/p}$, called the modified

Cook's D, can be judged in terms of a standard normal distribution, implying that we can use 2 or 3 as the cutoff value.

5. Empirical Illustrations

This section illustrates the performance of the proposed statistics in Section 4. We will use the 1998 Survey of Mental Health Organizations (SMHO) which contains a variety of variables that are suitable for linear regression analysis. The 1998 SMHO collected data on approximately 1,530 specialty mental health care organizations and general hospitals that provide mental health care services, with an objective to develop national and state level estimates for total expenditure, full-time equivalent staff, bed count, and total caseload by type of organization. The sample for this survey was based on a stratified single-stage design with probability proportional to size (PPS) sampling and is described in more detail in Li and Valliant (2009) and Manderscheid and Henderson (2002). The measure of size (MOS) used in sampling was the number of "episodes," defined as the number of patients/clients of an organization at the beginning of 1998 plus the number of new patients/clients added during calendar year 1998. The varying sizes of the mental health care organizations result in the values of collected variables in the sample having wide ranges, which may cause some observations to have relatively large influence on the parameter estimates of a linear regression.

The model of interest is to regress the total expenditure of a health organization on the number of beds set up and staffed for use and the number of additions of patients or clients during the reporting year. The total expenditure was defined as the sum of salary and contract personnel expenses, other contract and operating expenses, and depreciation expenses, and then divided by 1,000. The number of beds was the total number of hospital beds and residential beds. Scatterplots of expenditures versus beds and additions are shown in Figure 1. This figure will be described in more detail later in connection with the DFBETAS analysis. Note that there is one extreme value in the upper right of each panel. This case has an extremely large total expenditure ($519,863.3 in thousands of U.S. dollars), number of beds (2,405), and number of additions (79,808), but a small sample weight of 2.22. We omit it from several of the subsequent plots to avoid distorting their scales. It is detected as influential by all OLS and SW diagnostics.

For this illustration, we ignored the stratification in the original sampling design and treated the sample as selected using single-stage sampling with varying selection probabilities. Of course, the diagnostics given earlier will accommodate stratification through the use of appropriate variance formulae. A total of 875 observations was used in the regression due to missing values in the independent and dependent variables.

Table 1 gives a summary of the quantile values of the variables involved in the regression, including the survey weights. The total expenditure has a maximum of 519,863.3, which is almost 30,000 times the minimum, 16.6. The number of beds and the number of additions also have significant differences between their maxima and minima. Sample weights range from 1 to 158.86. The weights we use in analysis include a nonresponse adjustment which was done separately by design stratum. In some cases, units that were selected with certainty in the initial sample did not respond and some of the

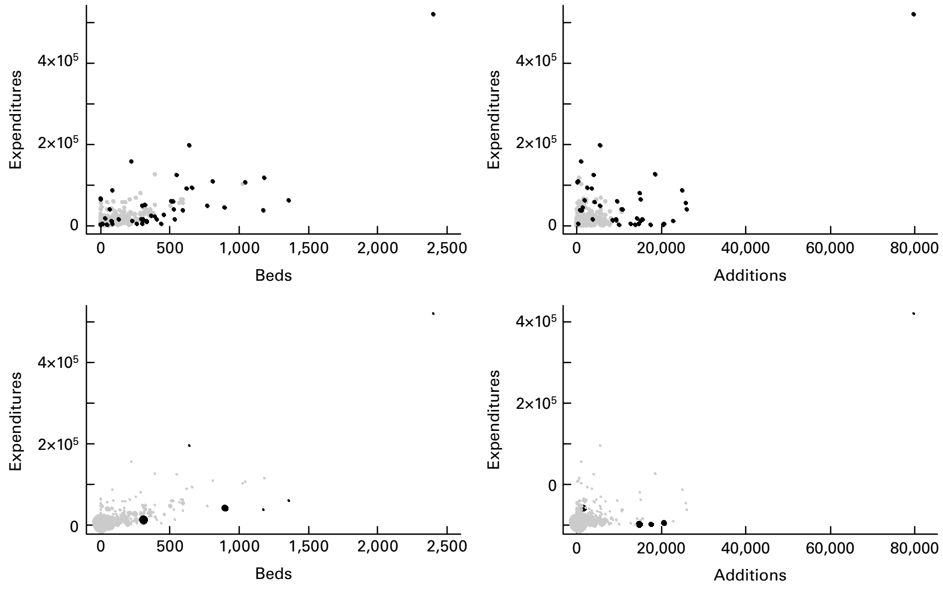

Fig. 1. Scatterplots of expenditures vs. beds and adds. In the top row, dark points are ones with OLS DFBETAS $> 3/\sqrt{n}$. In the bottom row, dark points are ones with SW DFBETAS $> 3/\sqrt{n}$

responding certainties had their weights adjusted to be larger than 1. A total of 157 organizations had a weight of 1 after the nonresponse adjustment.

In the SW regressions that follow, variances and standard errors were estimated using the sandwich estimator of Binder (1983), which is the type of linearization estimator included in most software packages that handle survey data. We have included the units with weights of 1 in standard error estimation rather than excluding them, as would be the approach for handling certainties in purely design-based estimation. Including the certainties is consistent with the idea that a superpopulation model is being estimated and that slope coefficients would still have a variance even if a census were done. A sketch of the mathematical justification for doing this is model-dependent (not design-based) and is given in the appendix of Li and Valliant (2009).

5.1. Parameter Estimation

The identification of single influential points will be compared under two different settings. One is to assume the sample design can be ignored and that the data can be

Table 1. Quantiles of Variables in SMHO Regression

	Quantiles				
Variables	0%	25%	50%	75%	100%
Expenditure (1,000's)	16.6	2,932.5	6,240.5	11,842.6	519,863.3
# of beds	0	6.5	36	93	2,405
# of additions	0	558.5	1,410	2,406	79,808
Weights	1	1.42	2.48	7.76	158.86

analyzed by conventional OLS regression estimators. The other is to account for the single-stage sampling design with varying survey weights and incorporate these weights into PML estimation. Which approach is preferable is a recurring question among analysts and is discussed at length in the collections edited by Skinner et al. (1989) and Chambers and Skinner (2003). Rather than recommending one approach over the other for this example, we will consider the possibility of there being two analysts, one who, after careful thought, chooses to use OLS and another who elects to use SW least squares. We illustrate how the parameter estimates and points identified as being influential can be different in OLS and SW regression.

The estimated coefficients and their standard errors are reported in Table 2. The intercept and slope coefficients all have discrepancies between the two methods, and the estimated intercept even changes from negative to positive and from significant with OLS to nonsignificant with SW. The relative size of the differences between the OLS and SW estimates is much greater for the intercept than the slopes. Analysts are often more focused on the latter. The *t*-statistics are also much smaller in the SW regression than in OLS owing to the substantially larger SW standard errors. (An alternative, not shown here, would be to use standard error (SE) estimates for the OLS parameter estimates that are robust to heteroscedasticity; e.g., see Long and Ervin (2000). The estimated OLS SE's would be somewhat larger with those estimates).

5.2. Leverages and Residuals

Figure 2, in the left-hand panel, shows a scatterplot of leverages calculated using the two methods with and without sample weights. The areas of the bubbles in this and later figures are proportional to the sample weights. Outlying points, with leverages greater than twice their mean, were identified to be the ones beyond the two reference lines. Twenty-seven outlying observations were identified by the SW but not by the OLS diagnostics and are represented by relatively large bubbles in area A. These points are associated with large sample weights ranging from 7.44 to 158.86; whereas the 14 outlying observations identified by OLS only, represented by small bubbles in area B, have small weights ranging from 1 to 2.62. The bubbles in the upper right square, with moderate sizes, stand for the points identified by both methods.

The points in the residual plot in the right-hand panel show the residuals scaled by the estimated standard error $\hat{\sigma}$ of model (1), where $\hat{\sigma}$ was estimated by the OLS estimator for the OLS scaled residuals and by the SW formula (3) for the SW scaled residuals. With a few exceptions, the weighted and unweighted diagnostics identified similar extreme

Table 2. OLS and SW Parameter Estimates of SMHO Regression of Expenditures on Beds and Additions

Independent	OLS estimation			SW estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	− 1,201.73	526.19	− 2.28	514.08	1,157.71	0.44
# of beds	94.16	3.03	31.08	81.23	13.14	6.18
# of additions	2.31	0.13	18.50	1.84	0.76	2.43

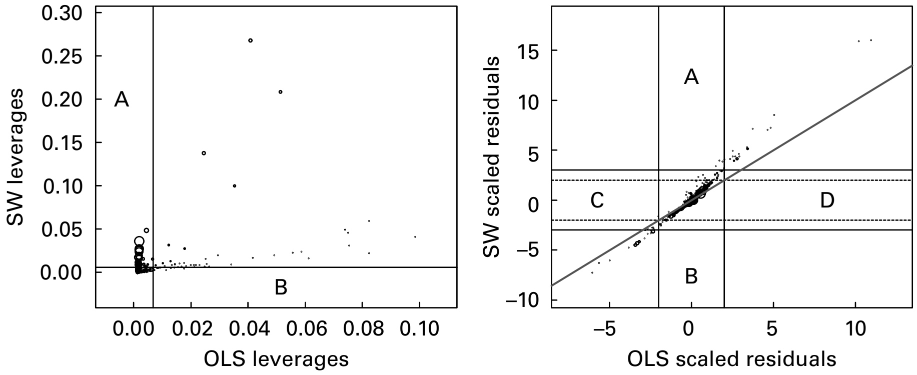

Fig. 2. Leverage and Residual Diagnostic Plots for SMHO Data. In the leverage plot on the left, area A includes points identified as outlying by SW diagnostic only; area B includes points identified by OLS diagnostic only. In the residual plot on the right, areas A and B include points identified by SW only; areas C and D include points identified by OLS only. A diagonal reference line is drawn at 45 degrees. Horizontal and vertical reference lines in the left-hand panel are drawn at 2p/n; reference lines in the right-hand panel are at $\pm 3/\sqrt{n}$ *(solid line) and* $\pm 2/\sqrt{n}$ *(dotted line)*

residuals. The residual analysis mainly filters out the observations with outlying Y values, but not necessarily those with outlying weights.

Table 3 lists the coefficient estimates on the reduced samples excluding the detected outlying observations. Qualitatively, the conclusion would be the same whether one uses OLS or SW diagnostics – all three parameter estimates are significantly different from zero. A more quantitative measure of difference is obtained by comparing predicted values calculated after excluding units identified by the OLS and SW diagnostics. Figure 3 displays the resultant fitted values versus those from the full sample. The slope coefficients decreased when the outliers were not used in the regressions, which accordingly resulted in smaller fitted values.

Table 3. OLS and SW Parameter Estimates after Deleting Observations with Large Leverages from SMHO Regression

	OLS estimation			SW estimation		
(i) Deleting units with leverages greater than $2p/n = 0.007$						
No. of units deleted	48			61		
Independent variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	2,987.55	490.54	6.09	1,993.86	353.71	5.64
# of beds	69.27	4.347	15.94	75.82	6.75	11.23
# of additions	0.947	0.201	4.71	0.997	0.211	4.73
(ii) Deleting units with absolute standardized residuals greater than 3						
No. of units deleted	17			37		
Independent variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	645.83	311.63	2.07	1,674.66	386.27	4.34
# of beds	84.48	1.98	42.67	76.19	5.28	14.43
# of additions	1.531	0.103	14.86	0.932	0.217	4.29

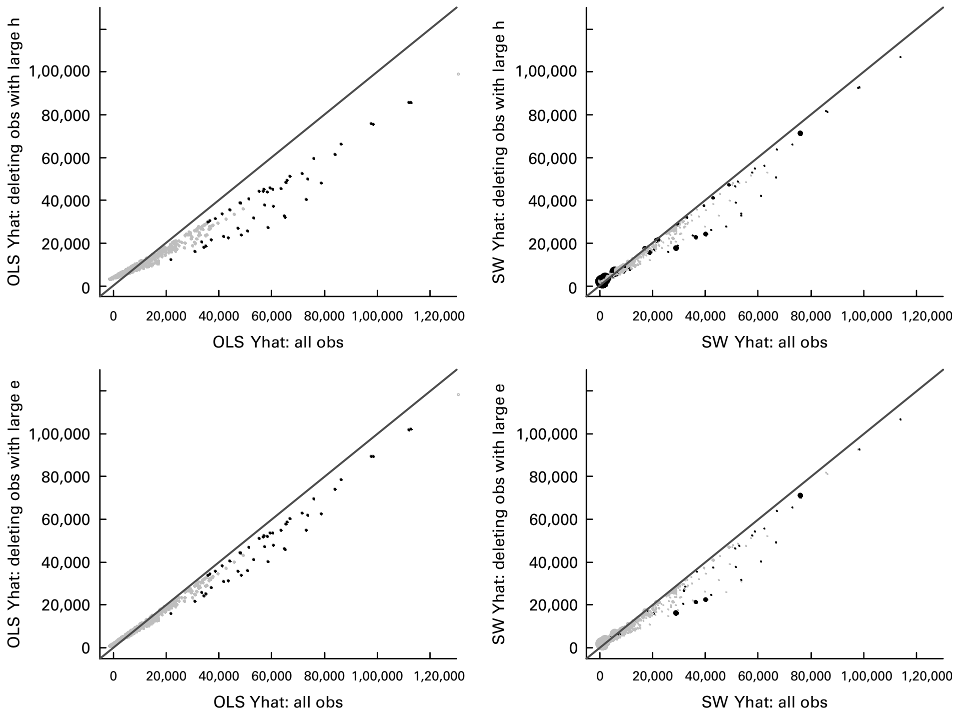

Fig. 3. Fitted Values Plots After Applying Leverage and Residual Diagnostics to SMHO Data. The first row plots the fitted values from the regression after deleting observations with leverages greater than 2p/n versus those from the regression on full sample for both OLS and SW. The second row is based on deleting units with standardized residuals greater than 3. Points in grey are ones not identified by the diagnostics; points in black are ones identified as influential. A 45 degree line is drawn in each panel

5.3. *DFBETAS*

The scatterplots in Figure 1 show the points in bold in the first row having OLS DFBETAS greater than $3/\sqrt{n}$; the bold points in the second row are ones with SW DFBETAS greater than $3/\sqrt{n}$. In the second row the sizes of points are proportional to the survey weights. The diagnostic results for the DFBETAS statistics for number of beds and number of additions are also graphically presented in Figure 4. The figure clearly shows, especially in the partially enlarged graphs at the second row, that points identified only by the OLS method have small weights symbolized by the bubbles of small sizes. The SW DFBETAS singled out a few points associated with moderate sampling weights, most of which were also identified by OLS.

Table 4 reports the estimated coefficients and their standard errors when the outliers identified by DFBETAS (both DFBETAS for beds and DFBETAS for additions are beyond the cutoff) were removed from the sample. OLS flags many more cases – 57 compared to 9 – than does SW. After deleting the 57 cases, the OLS intercept reverses sign from $-1{,}201.73$ to 2,044.54. The slope for beds changes somewhat but the slope for adds goes from 2.31 to 0.96 (compare Tables 2 and 4). On excluding the observations with large SW DFBETAS for either number of beds or adds, the slope for

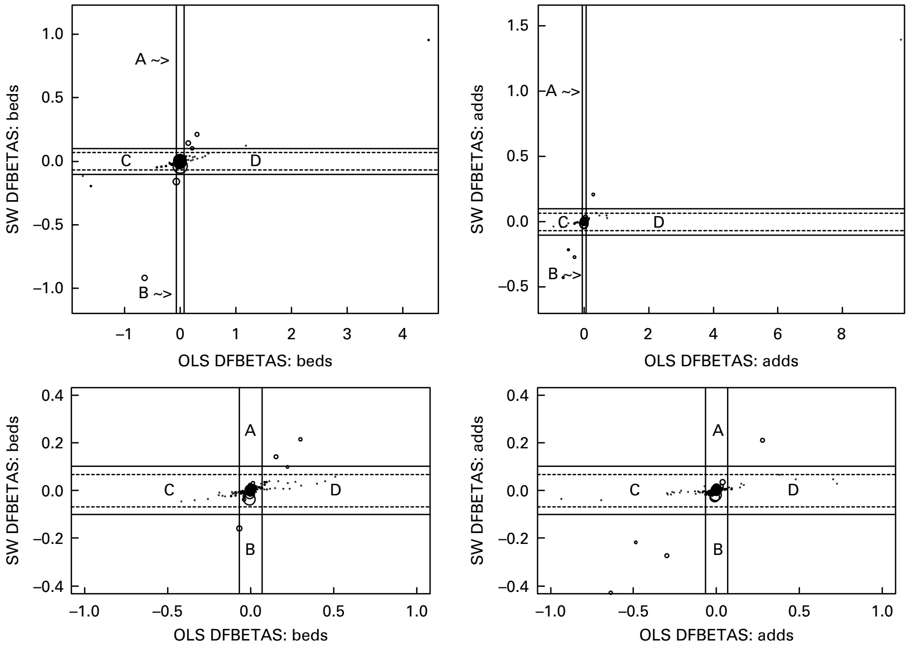

Fig. 4. DFBETAS Plots for SMHO Data. Areas A and B include points identified only by the SW diagnostics whereas areas C and D include points identified by the OLS diagnostics only. Partially enlarged graphs are presented in the second row. Solid reference lines are drawn at ±$3/\sqrt{n}$; dotted at ±$2/\sqrt{n}$

beds changes very little but the standard error drops from 13.14 to 4.49 (compare Tables 2 and 4). The slope for adds drops from 1.84 to 1.27 and the standard error reduces from 0.76 to 0.28.

The OLS diagnostics identified many more points as being influential than the SW diagnostics did. We also analyzed predictions but do not show the details here. Dropping points flagged as influential by DFBETAS leads to systematic reductions in predicted values for OLS predictions when these points are omitted. The SW analysis omits fewer points and has less of an effect on predictions. Thus, if an analyst takes the position that the sample design is ignorable, does not use weights, and applies OLS diagnostics, substantially different predictions will be obtained in this case.

Table 4. OLS and SW Parameter Estimates after Deleting Observations with DFBETAS for beds and adds > $3/\sqrt{n}$

	OLS estimation			SW estimation		
No. units deleted	57			9		
Independent variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	2,044.54	353.01	5.79	1,485.03	425.83	3.49
# of beds	82.36	2.61	31.55	81.72	4.49	18.19
# of additions	0.96	0.15	6.42	1.27	0.28	4.59

5.4. DFFITS and Modified Cook's Distance

Both DFFITS and modified Cook's Distance statistics summarize the effect of deleting a specific unit on the overall parameter estimation. Figure 5 plots the OLS and SW versions of DFFITS and the modified Cook's D versus each other. Tables 5 and 6 show the parameter estimates after deleting the points identified as influential by the two methods. There were nine units identified as influential by SW DFFITS; of these three were flagged only by SW. The range of the weights for cases identified by SW but not by the OLS diagnostics was 37.8 to 158.86. Forty-five cases were identified as influential by OLS DFFITS; of these 39 were found by OLS not SW. Their weights were relatively small, ranging from 1, which is the smallest weight in the sample, to 5.5. The SW modified Cook's Distance identified 10 cases (of which four were not flagged by OLS), with weights from 11.38 to 158.86. The OLS Cook's Distance detected 44 points, of which 38 points were not found by SW. The 38 cases also had weights that ranged from 1 to 5.5. No cases with large weights were identified by the OLS Cook's Distance.

There was only one observation identified by the OLS modified Cook's Distance but not by the OLS DFFITS. As a result, the parameter estimates based on the samples without

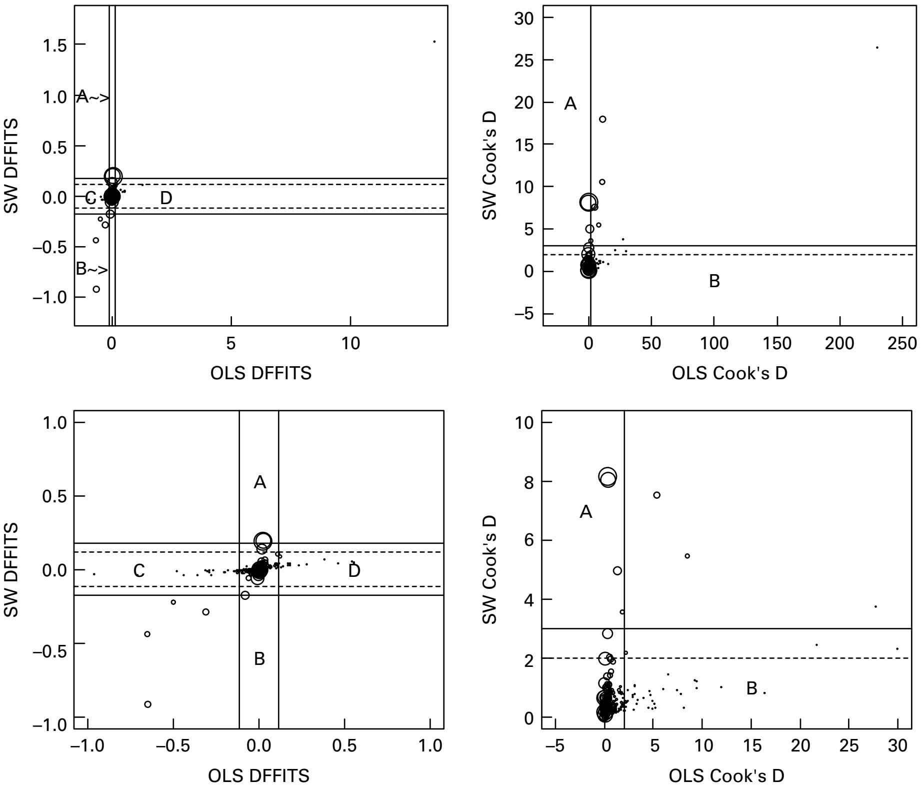

Fig. 5. DFFITS and Modified Cook's Distance Plots for SMHO Data. Areas A and B in the DFFITS plot and area A in the Cook's Distance plot include points identified only by the SW diagnostics; areas C and D in the DFFITS plot and area B in the Cook's Distance plot include points identified by the OLS diagnostics only. Partially enlarged graphs are presented below the originals

Table 5. OLS and SW Parameter Estimates after Deleting Observations with Large DFFITS. 39 units were identified by OLS only, 3 units by SW only

	OLS estimation			SW estimation		
No. units deleted	45			9		
Independent variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	1,617.67	335.38	4.82	1,028.71	360.46	2.85
# of beds	81.45	2.44	33.38	82.94	5.72	14.50
# of additions	1.20	0.12	9.77	1.40	0.27	5.27

Table 6. OLS and SW Parameter Estimates after Deleting Observations with Large Modified Cook's Distance. 38 units were identified by OLS only, 4 units by SW only

	OLS estimation			SW estimation		
No. units deleted	44			10		
Independent variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	1,660.45	335.54	4.95	932.43	345.86	2.70
# of beds	80.92	2.44	33.16	82.83	5.72	14.48
# of additions	1.19	0.12	9.66	1.43	0.26	5.43

the identified outliers are very similar for these two cases (compare Tables 5 and 6). The estimated slopes dropped moderately compared to the ones from the full sample in Table 2. For the SW diagnostics, the two statistics also have comparable performance. Since fewer outliers were picked from the sample by the SW DFFITS and the SW modified Cook's Distance, the SW estimates from the reduced samples changed less than the OLS ones. Comparing to Table 2, we see that the standard errors again decrease substantially after deleting cases, particularly for SW.

6. Conclusion

The conventional OLS influence diagnostics require modification to be used for complex sample data to accommodate survey weights and variances. We take the point of view that the goal of model-fitting is to identify a model that fits for the bulk of the points in a finite population. Thus, the goal of using diagnostics is to detect points that do not follow that core model. Some points in a sample can be influential in the sense that the parameter estimates may change noticeably if they are dropped from the fitting. With survey data the influence may be due to extreme values of the response variable, the covariates, or the survey weights.

As for many of the conventional diagnostics, little formal distribution theory is available, and heuristic arguments must be used to identify points that are influential in model fitting. The cutoff values for the adapted statistics presented here were determined and justified in terms of model distributions and various order of magnitude arguments.

Based on the comparison of the OLS and the SW influence analysis on a sample of mental health organizations, we conclude that the SW diagnostics, including leverages, residuals, DFBETAS, DFFITS, and a modified Cook's Distance, can identify a different set of points than the OLS diagnostics as being influential. Different diagnostic statistics identify different sets because they focus on measuring different kinds of changes in the regression estimation after a point is deleted from the sample. Even in a large sample, dropping a few influential points can have a substantial effect on estimates, particularly standard errors, in survey weighted regressions.

Note that there can be situations where points with large weights, residuals, or **X** values would be important in identifying whether a model is correctly specified. For example, if Y were quadratically related to an x and units with large **X**'s were deleted because of large weights or large residuals, the ability could be lost to recognize that the model should be quadratic. An analyst must make a decision on whether to include or exclude a point based on what he/she knows about the substantive problem and the way that the survey data were collected and processed. In general, diagnostics should be treated as exploratory tools that should be applied with care.

Areas for additional research are the identification of groups of influential points and the adaptation of diagnostics to situations where models involving clustering are appropriate. The diagnostics presented here depend on variance estimates for regression parameters that assume no clustering. Modifications of the statistics and the arguments justifying heuristic cutoffs are needed to account for clustering. We plan to report results on these topics in future research.

7. References

Atkinson, A.C. (1982). Regression Diagnostics, Transformations and Constructed Variables (with Discussion). Journal of the Royal Statistical Society, Series B, Methodological, 44, 1–36.

Belsley, D.A., Kuh, E., and Welsch, R. (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.

Brewer, K.R.W. and Särndal, C.-E. (1983). Six Approaches to Enumerative Survey Sampling. In Incomplete Data in Sample Surveys, W.G. Madow and I. Olkin (eds). Vol. 3. New York: Academic Press, 363–368.

Chambers, R.L. (1986). Outlier Robust Finite Population Estimation. Journal of the American Statistical Association, 81, 1063–1069.

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

Cook, R.D. (1977). Detection of Influential Observation in Linear Regression. Technometrics, 19, 15–18.

Cook, R.D. and Weisberg, S. (1982). Residuals and Influence in Regression. London: Chapman & Hall Ltd.

DuMouchel, W.H. and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association, 78, 535–543.

Ericksen, E.P. (1988). Estimating the Concentration of Wealth in America. Public Opinion Quarterly, 52, 243–253.

Gambino, J. (1987). Dealing with Outliers: A Look at Some Methods used at Statistics Canada. Paper prepared for the Fifth Meetings of the Advisory Committee on Statistical Methods, Ottawa: Statistics Canada.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous Testing of Regression Coefficients with Complex Survey Data: Use of Bonferroni Statistics. The American Statistician, 44, 270–276.

Korn, E.L. and Graubard, B.I. (1995). Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. The American Statistician, 49, 291–295.

Kott, P.S. (1991). A Model-based Look at Linear Regression with Survey Data. The American Statistician, 45, 107–112.

Lee, H. (1995). Outliers in Business Surveys. Chapter 26 in Business Survey Methods, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (eds). New York: John Wiley.

Li, J. (2007). Regression Diagnostics for Complex Survey Data: Identification of Influential Observations. Unpublished doctoral dissertation, University of Maryland.

Li, J. and Valliant, R. (2009). Survey Weighted Hat Matrix and Leverages. Survey Methodology, 35, 15–24.

Little, R.J.A. (1991). Inference with Survey Weights. Journal of Official Statistics, 7, 405–424.

Long, J.S. and Ervin, L.H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. The American Statistician, 54, 217–224.

Manderscheid, R.W. and Henderson, M.J. (2002). Mental Health, United States, 2002. DHHS Publication No. SMA04-3938. Rockville MD USA: Substance Abuse and Mental Health Services Administration. Available at http://mentalhealth.samhsa.gov/publications/allpubs/SMA04-3938/AppendixA.asp

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). Applied Linear Statistical Models, (Fourth edition). Homewood, IL: Richard D. Irwin Inc.

Pfeffermann, D. and Holmes, D.J. (1985). Robustness Considerations in the Choice of Method of Inference for the Regression Analysis of Survey Data. Journal of the Royal Statistical Society, Series A, 148, 268–278.

Pukelsheim, F. (1994). The Three Sigma Rule. The American Statistician, 48, 88–91.

Rubin, D.B. (1985). The Use of Propensity Scores in Applied Bayesian Inference. In Bayesian Statistics 2, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds). Amsterdam: North Holland.

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

Smith, T.M.F. (1987). Influential Observations in Survey Sampling. Journal of Applied Statistics, 14, 143–152.

Smith, T.M.F. (1988). To Weight or Not To Weight: That Is The Question. In Bayesian Statistics 3, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds). Oxford, UK: Oxford University.

Srinath, K.P. (1987). Outliers in Sample Surveys. Paper prepared for the Fifth Meetings of the Advisory Committee on Statistical Methods, Ottawa: Statistics Canada.

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

Weisberg, S. (2005). Applied Linear Regression, (Third Edition). New York: John Wiley.

Received July 2009
Revised December 2010