Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Survey weighted hat matrix and leverages

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

Canada

by Jianzhu Li and Richard Valliant

June 2009

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 15-24
Statistics Canada, Catalogue No. 12-001-X

Survey weighted hat matrix and leverages

Jianzhu Li and Richard Valliant [1]

Abstract

Regression diagnostics are geared toward identifying individual points or groups of points that have an important influence on a fitted model. When fitting a model with survey data, the sources of influence are the response variable **Y**, the predictor variables **X**, and the survey weights, **W**. This article discusses the use of the hat matrix and leverages to identify points that may be influential in fitting linear models due to large weights or values of predictors. We also contrast findings that an analyst will obtain if ordinary least squares is used rather than survey weighted least squares to determine which points are influential.

Key Words: Influence; Linear regression; Survey data; Weighted least squares.

1. Introduction

In some conventional linear regression diagnostics, it is often useful to measure the influence each data point can have in determining the values of parameter estimates and, in turn, fitted values. The hat matrix and its diagonal elements, referred to as leverages, are popular techniques that are used to identify the cases that have outlying values for predictor variables, and, therefore, may be influential in model fitting if they are also associated with unusual residuals. When there is more than one predictor variable in the regression, analysts can compute leverages to summarize the collective influence of the **X** values for each observation.

In finite population estimation, a superpopulation assumption is usually used to build models. Suppose that some model fits reasonably well for the bulk of the population. For convenience, we will refer to this as the "true" model. However, the goal is usually to find a model that has some descriptive or predictive power, bearing in mind that no model is really "true". The influence diagnostics should allow analysts to identify points that make estimated parameters deviate from that true model. Parameter estimates in linear regression using complex survey data are often derived from the pseudo maximum likelihood approach, outlined by Skinner, Holt and Smith (1989, Chapter 3), following ideas of Binder (1983). In this paper, we assume that the analyst has decided that an estimator involving sample weights is appropriate for his or her problem. As shown in later sections, the survey weighted hat matrix and leverages are useful for detecting potentially influential observations caused by not only extreme **X** values, but also by large sample weights.

Previous survey literature has discussed the effect of outliers on some survey estimates, but does not give much attention to diagnostics for linear regression models. Deville and Särndal (1992), and Potter (1990, 1993) discuss some possibilities for locating or trimming extreme survey weights when the goal is to estimate population totals and other simple descriptive statistics. Hulliger (1995) and Moreno-Rebollo, Muñoz-Reyes and Muñoz-Pichardo (1999) address the effect of outliers on the Horvitz-Thompson estimator of a population total. Smith (1987) demonstrates diagnostics based on case deletion and a form of the influence function. Zaslavsky, Schenker and Belin (2001), and Beaumont and Alavi (2004) use M-estimation based strategies to downweight the influential clusters or units. Chambers (1986), Gwet and Rivest (1992), Welsh and Ronchetti (1998), and Duchesne (1999) conduct research on outlier robust estimation techniques for totals.

A perennial question among analysts of survey data is whether to use the survey weights or not when fitting models. The collections edited by Skinner *et al.* (1989) and Chambers and Skinner (2003) discuss this issue at length. Binder and Roberts (2003, Chapter 3), Chambers, Dorfman and Sverchkov (2003, Sections 11.2.3, 11.6), Chambers and Skinner (2003, Chapter 1), Korn and Graubard (1999, Sections 4.3, 4.4), Pfeffermann (1996), and Smith (1989, Chapter 6) describe the arguments pro and con. The details can be quite mathematical and abstract but are summarized succinctly by Skinner (2003, Section 6.2.3).

We paraphrase Skinner (2003, Section 6.2.3) here in the context of fitting a linear model to predict some response Y based on a set of explanatory variables **X**. If the linear model is specified correctly and the sampling depends only on the explanatory variables in the model, then unweighted regression parameter estimates will be unbiased in a model-based sense. In particular, the assumed conditions require that the survey weights are unrelated to Y conditional on the values of the **X** predictors. However, if sampling depends on factors that may be related to Y, even after conditioning on the values of the predictors, the unweighted parameter

[1] Jianzhu Li, Westat, 1650 Research Boulevard, Rockville MD 20850; Richard Valliant, Survey Research Center, University of Michigan, and Joint Program in Survey Methodology, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742.

estimators will be biased both with respect to the true model and in the design-based, repeated sampling sense. This situation is known as having an *informative* sample design in which the distribution of the sample values of Y is different from the population distribution. An example of this is given by Chambers, Dorfman and Sverchkov (2003, Section 11.2.3). If sample units are selected with probabilities proportional to some measure x of their size and Y is related to x, the sample distribution of Y will be skewed to the right of its population distribution. The situation in this example is similar to the one in our empirical study in section 5.

Using the survey weights guards against the bias that may result from not accounting for an informative sample. Also, if the model is not correctly specified, the survey-weighted regression still estimates a census parameter. That is, the weighted estimates are approximately unbiased for the best-fitting linear model that would be obtained if the entire finite population were in hand. In this paper, we assume that an analyst has made the decision to use weights in fitting a model, possibly for the reasons above, and provide one type of diagnostic for assessing the effects of certain data points.

The hat matrix and leverages we present are the same ones that are produced by standard software packages when a weighted least squares regression is done. However, the literature is missing any discussion of their use and interpretation in the context of survey-weighted regression. Korn and Graubard (1999) is one of the few references that addresses any kind of diagnostics for models fitted from survey data. Leverages are among a series of diagnostic tools and will be more effective when evaluated with residuals. Many diagnostic statistics, such as the famous Cook's distance (Cook 1977) turn out to have both leverages and residuals as components.

The literature gives somewhat ambiguous guidance on how to deal with the influential observations once they are identified. An obvious, and perhaps naïve, solution is to remove the outliers and refit the model, which makes sense when the outliers result from improperly recorded data. A natural extension of this would be to devise an automatic approach where certain rules would be used to identify influential points, delete them, and refit the model. Our presumption in this article is that, after identification of influential points and careful consideration of the reasons for the influence, an analyst will determine whether the points should be excluded from fitting. This is in contrast to setting up some procedure that would automatically exclude points based on some cutoff values.

The remainder of the paper is organized as follows. Section 2 describes the ordinary least squares hat matrix, leverages, and some of their properties. Sections 3 and 4 cover the survey-weighted hat matrix and leverages plus a decomposition that shows how points can have large leverages. The extensions to survey data apply to both single- and multi-stage designs. Section 5 gives a numerical example using a single-stage sample of mental health organizations. The last section summarizes our findings and gives some directions for additional research.

2. OLS hat matrix

A *working* model is one that is being provisionally considered by an analyst for the structure that best describes a conceptual superpopulation. It may be revised after further assessment by adding predictors, dropping predictors, or making other changes to the form of the model. Suppose that the working linear model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I} \tag{1}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)^T$, $\mathbf{X}^T = (\mathbf{x}_1, ..., \mathbf{x}_n)$ with $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$. Assuming the $\mathbf{X}$ matrix is of full rank, the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2}$$

where $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ is a square matrix and invertible. The fitted values $\hat{\mathbf{Y}}$ corresponding to the observed values $\mathbf{Y}$ are

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T$ is called the hat matrix. This name was first introduced by Tukey (Belsley, Kuh and Welsch 1980, Chapter 2; Hoaglin and Welsch 1978). The leverage, $h_{ii} = \mathbf{x}_i^T\mathbf{A}^{-1}\mathbf{x}_i$, is the i^{th} element on the diagonal of the hat matrix, which measures the impact of Y_i on its own fitted value since $\hat{Y}_i = \sum_j h_{ij}Y_j = h_{ii}Y_i + \sum_{j\neq i} h_{ij}Y_j$. If h_{ii} approaches 1, Y_i has a crucial role in determining the value of $\hat{Y}_i$.

The OLS hat matrix and leverages have many special and useful properties:

(i) $\mathbf{H}$ is symmetric, or $h_{ij} = h_{ji}$;

(ii) $\mathbf{H}$ is idempotent, or $\mathbf{H} = \mathbf{H}^2$, or $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$;

(iii) $\mathbf{H}\mathbf{X} = \mathbf{X}$ or $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$;

(iv) $0 \le h_{ii} \le 1$;

(v) $\sum_i h_{ii} = \text{rank}(\mathbf{X}) = p$, which implies that the mean leverage is $\bar{h} = p/n$;

if model (1) has an intercept, the following two properties hold:

(vi) $\sum_i h_{ij} = 1$;

(vii) $h_{ii} = 1/n + (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{A}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}} = \sum_n \mathbf{x}_i / n$.

In a reasonably large data set, an individual leverage value h_{ii} is usually considered extreme if it is more than twice the mean, $\bar{h} = p/n$ (Belsley *et al.* 1980, Chapter 2). The existence of a gap between most of the cases and a few unusual cases in the empirical distribution of the leverages also provides evidence of outlying units.

3. Survey weighted hat matrix

The initial step in the pseudo maximum likelihood approach is to form the set of estimating equations that would be appropriate for a model if the entire finite population were observed. This set is a type of population total which is then estimated using design-based survey methods. Suppose that the underlying structural model is a fixed-effects linear model:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \text{ind}\, N(0, v_i \sigma^2) \tag{3}$$

where ε_i is independently normally distributed with mean 0 and variance $v_i \sigma^2$, which is known except for the constant σ^2. The pseudo maximum likelihood estimator (PMLE) of $\boldsymbol{\beta}$ is the solution to the set of estimating equations $\mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$, with $\mathbf{V} = \text{diag}(v_1, ..., v_n)$ and $\mathbf{W} = \text{diag}(w_1, ..., w_n)$. Survey weights, which in probability samples are usually inversely proportional to inclusion probabilities, are used in the PMLE to account for an informative design in which the sample distribution of the Y's is likely to differ from that of the finite population. These equations can be solved explicitly as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{Y}$. If we assume $\mathbf{V} = \mathbf{I}$, model (3) reduces to (1) and the survey-weighted (SW) estimator $\hat{\boldsymbol{\beta}}$ will consequently take the form of a weighted least squares estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$.

When survey weights are accounted for in the regression, the predicted values become $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$, where the hat matrix includes the survey weights and is defined as

$$\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} = \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W}$$

with $\mathbf{A} = \mathbf{X}^T \mathbf{W} \mathbf{X}$. The leverages on the diagonal of the hat matrix are $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i w_i$. In this formulation, it is assumed that the analyst does not incorporate a $\mathbf{V}$ matrix in the regression. However, results can be modified to incorporate $\mathbf{V}$ simply by using $\mathbf{W}^* = \mathbf{W}\mathbf{V}^{-1}$ rather than $\mathbf{W}$. Unlike the unweighted hat matrix, the SW hat matrix is no longer symmetric for sampling designs with unequal selection probabilities (or, more generally, unequal weights). Properties (ii) – (vi) in section 2 still hold (*e.g.*, see Valliant, Dorfman and Royall 2000, Chapter 5) provided the unweighted hat matrices were replaced by the weighted ones. In addition, the SW hat matrix has extra useful, and easily verified, properties as follows:

a) $\mathbf{W}\mathbf{H} = \mathbf{W}\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T \mathbf{W} = \mathbf{H}^T \mathbf{W}$;

b) $\mathbf{X}^T \mathbf{W} (\mathbf{I} - \mathbf{H}) = \mathbf{X}^T \mathbf{W} - \mathbf{X}^T \mathbf{H}^T \mathbf{W} = \mathbf{0}$;

c) $w_{i'} h_{i'i} = w_{i'} \mathbf{x}_{i'}^T \mathbf{A}^{-1} \mathbf{x}_i w_i = w_i h_{ii'}$.

The definition of the weighted leverages indicates that a large leverage may be caused by outlying $\mathbf{X}$ values, an outlying weight, or both. Note that the formulas for the survey-weighted hat matrix and leverages apply regardless of whether the sample design uses strata or is single-stage or multi-stage. This is in contrast to diagnostics, like Cook's D, that require estimated standard errors or covariance matrices that should be specialized to fit the sample design.

4. Decomposition of leverages

Leverages can be decomposed into components that separate the effect of the weight and the $\mathbf{X}$ values for a unit. Suppose the working model is (1) and that the model contains an intercept, so that

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^T \end{pmatrix} \equiv (\mathbf{1}\ \mathbf{X}_1), \text{ and } \mathbf{X}_1 = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix},$$

where $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{i,p-1})$ are $1 \times (p-1)$ vectors, $\mathbf{1}$ is a $n \times 1$ vector with all the elements equal to 1, and $\mathbf{X}_1$ is a $n \times (p-1)$ matrix. The $\mathbf{A}$ matrix is computed as

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}^T \\ \mathbf{X}_1^T \end{pmatrix} \mathbf{W} (\mathbf{1}\ \mathbf{X}_1) = \begin{pmatrix} \mathbf{1}^T \mathbf{W} \mathbf{1} & \mathbf{1}^T \mathbf{W} \mathbf{X}_1 \\ \mathbf{X}_1^T \mathbf{W} \mathbf{1} & \mathbf{X}_1^T \mathbf{W} \mathbf{X}_1 \end{pmatrix} \equiv \begin{pmatrix} \hat{N} & \hat{\mathbf{t}}_X^T \\ \hat{\mathbf{t}}_X & \mathbf{A}_1 \end{pmatrix},$$

where $\hat{\mathbf{t}}_X$ is a $(p-1) \times 1$ vector with elements $\hat{\mathbf{t}}_{Xj} = \sum_{i \in s} w_i x_{ij}$ and $\mathbf{A}_1$ is a $(p-1) \times (p-1)$ matrix. Using the inverse of a partitioned matrix,

$$\mathbf{A}^{-1} = \begin{pmatrix} \dfrac{1}{\hat{N}} + \dfrac{1}{\hat{N}} \hat{\mathbf{t}}_X^T \mathbf{S}^{-1} \hat{\mathbf{t}}_X \dfrac{1}{\hat{N}} & -\dfrac{1}{\hat{N}} \hat{\mathbf{t}}_X^T \mathbf{S}^{-1} \\ -\dfrac{1}{\hat{N}} \mathbf{S}^{-1} \hat{\mathbf{t}}_X & \mathbf{S}^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{1}{\hat{N}} + \bar{\mathbf{x}}_W^T \mathbf{S}^{-1} \bar{\mathbf{x}}_W & -\bar{\mathbf{x}}_W^T \mathbf{S}^{-1} \\ -\mathbf{S}^{-1} \bar{\mathbf{x}}_W & \mathbf{S}^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{1}{\hat{N}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\bar{\mathbf{x}}_W^T \\ \mathbf{I} \end{pmatrix} \mathbf{S}^{-1} (-\bar{\mathbf{x}}_W\ \mathbf{I})$$

where $\bar{\mathbf{x}}_W = \hat{\mathbf{t}}_X / \hat{N}$ is a $(p-1)\times 1$ vector, and $\mathbf{S} = \mathbf{A}_1 - \hat{\mathbf{t}}_X \hat{\mathbf{t}}_X^T / \hat{N}$ is a $(p-1)\times(p-1)$ matrix. Simplifying the hat matrix using the above inverse matrix, we obtain

$$\begin{aligned}\mathbf{H} &= \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\,\mathbf{W}\\ &= \left\{\frac{1}{\hat{N}}\mathbf{1}\mathbf{1}^T + (\mathbf{X}_1 - \mathbf{1}\bar{\mathbf{x}}_W^T)\,\mathbf{S}^{-1}(-\bar{\mathbf{x}}_W\mathbf{1}^T + \mathbf{X}_1^T)\right\}\mathbf{W}\\ &= \left\{\frac{1}{\hat{N}}\mathbf{1}\mathbf{1}^T + \begin{pmatrix}\mathbf{x}_1^T - \bar{\mathbf{x}}_W^T\\ \vdots\\ \mathbf{x}_n^T - \bar{\mathbf{x}}_W^T\end{pmatrix}\mathbf{S}^{-1}(\mathbf{x}_1 - \bar{\mathbf{x}}_W, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}_W)\right\}\mathbf{W}.\end{aligned}$$

Then, using the fact that $\hat{N} = n\bar{w}$ with $\bar{w} = \sum_{i=1}^n w_i / n$, the leverage of i^{th} observation, or the i^{th} diagonal element of the weighted hat matrix $\mathbf{H}$, is

$$h_{ii} = \frac{1}{n}\frac{w_i}{\bar{w}}[1 + \hat{N}(\mathbf{x}_i - \bar{\mathbf{x}}_W)^T\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_W)].$$

The quadratic form, $(\mathbf{x}_i - \bar{\mathbf{x}}_W)^T\,\mathbf{S}^{-1}\,(\mathbf{x}_i - \bar{\mathbf{x}}_W)$, defines an ellipsoid centered at $\bar{\mathbf{x}}_W$ (*e.g.*, see Weisberg 2005, Chapter 8), and $\hat{N}(\mathbf{x}_i - \bar{\mathbf{x}}_W)^T\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_W)$ is the Mahalanobis distance from $\mathbf{x}_i$ to $\bar{\mathbf{x}}_W$. Consequently, a leverage can be large if (1) w_i is large, especially relative to the average weight $\bar{w}$; or (2) $\mathbf{x}_i$ is far from the weighted average, $\bar{\mathbf{x}}_W$, of the $\mathbf{X}$, in the metric determined by the matrix $\mathbf{S}$.

For example, in a simple linear model with only one auxiliary variable, $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim (0, \sigma^2)$, the leverage of the i^{th} observation is

$$h_{ii}^W = \frac{1}{n}\frac{w_i}{\bar{w}}\left[1 + \hat{N}\frac{(x_i - \bar{x}_W)^2}{\sum_{j=1}^n w_j\,(x_j - \bar{x}_W)^2}\right].$$

where $\bar{x}_W = \sum_i w_i\,x_i / \hat{N}$.

If the error terms in the model have a general variance structure $\boldsymbol{\varepsilon} \sim (0, \mathbf{V})$ and $\mathbf{V}$ is known, the hat matrix is then defined as $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\,\mathbf{X}^T\,\mathbf{W}\mathbf{V}^{-1}$ with

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix}\mathbf{1}^T\,\mathbf{W}\mathbf{V}^{-1}\mathbf{1} & \mathbf{1}^T\,\mathbf{W}\mathbf{V}^{-1}\,\mathbf{X}_1\\ \mathbf{X}_1^T\,\mathbf{V}^{-1}\mathbf{W}\mathbf{1} & \mathbf{X}_1^T\,\mathbf{W}\mathbf{V}^{-1}\,\mathbf{X}_1\end{pmatrix}\\ &= \begin{pmatrix}\sum_s w_i / v_i & \sum_s w_i\mathbf{x}_i^T / v_i\\ \sum_s w_i\mathbf{x}_i / v_i & \sum_s w_i\mathbf{x}_i\mathbf{x}_i^T / v_i\end{pmatrix}.\end{aligned}$$

A formula for $\mathbf{A}^{-1}$ like the one above applies with $\hat{\mathbf{t}}_{XV} = \sum_s w_i\,\mathbf{x}_i / v_i$, $\hat{N}_V = \sum_s w_i / v_i$, and $\mathbf{S}_V = \mathbf{X}_1^T\,\mathbf{W}\mathbf{V}^{-1}\mathbf{X}_1 - \hat{\mathbf{t}}_{XV}\,\hat{\mathbf{t}}_{XV}^T / \hat{N}_V$. If a general $\mathbf{V}$ is used, $\hat{\mathbf{t}}_{XV}$ and $\hat{N}_V$ no longer are design-based estimates of $\mathbf{T}_X$ and N but are estimates of $\mathbf{T}_{XV} = \sum_1^N \mathbf{x}_i / v_i$ and $N_V = \sum_1^N 1 / v_i$. The leverage of the i^{th} observation under this general model is

$$h_{ii} = \frac{w_i}{v_i\,\hat{N}_V}\,[1 + \hat{N}_V\,(\mathbf{x}_i - \bar{\mathbf{x}}_{WV})^T\,\mathbf{S}_V^{-1}\,(\mathbf{x}_i - \bar{\mathbf{x}}_{WV})].$$

5. Numerical example

As noted in section 1, arguments can be advanced to justify ignoring sample design features, generally, and weights, in particular, when fitting models. Roughly speaking, when a model conditions on all the design variables determining the sampling scheme and the model is correct for both the population and the sample, OLS regression can be used. Analysts may object to including design variables in a model because some are not scientifically interesting as predictors. In addition, conditioning on all design variables may not be possible, especially when the "sampling scheme" includes uncontrolled nonresponse that itself may be related to the response variable. As noted in section 1, SW provides a modicum of protection against having a misspecified model when the distribution of the sample Y's is different from that of the population due to the type of sample design used. Nevertheless, some analysts will contend that the sample design and survey weights can be ignored in specific applications and that OLS is appropriate. Thus, it is interesting to see how different the OLS diagnostics are from SW diagnostics in a real application. However, given a course of action, an analyst should use diagnostics consistent with the method of fitting. If OLS is used, the standard OLS diagnostics should be examined; if SW regression is used, SW diagnostics are appropriate. It may well be that different points are influential depending on whether one uses OLS or SW regression.

In this section we examine the hat matrix and leverages in a regression example using the 1998 Survey of Mental Health Organizations (SMHO) conducted in the U.S., which collected data on specialty mental health care organizations and general hospital mental health care services. The sample for this survey was based on a stratified single-stage design with probability proportional to size (PPS) sampling (Manderscheid and Henderson 2002; Choudhry 2000). 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. Many of the analysis variables in the survey are related to the MOS, and their unweighted sample distributions will be different from the population distributions since the sample tends to have larger size units. Thus, this design is potentially informative as defined in Chambers and Skinner (2003).

The varying sizes of the mental health care organizations resulted 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 in this study is to regress the total expenditure of a health organization, in 1,000's of dollars, 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 SW estimator, $\hat{\mathbf{\beta}} = (\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{WY}$, was used. Mimicking the procedure employed by most analysts, we did not incorporate a model variance matrix $\mathbf{V}$ in the estimate of the regression parameter. A total of 875 observations was used in the regression, each of which had non-missing values on 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. Although not as extreme as the total expenditure, the number of beds and the number of additions also have significant differences between their maxima and minima. Because the sample was selected using a PPS design, the sample weights were associated with the sizes of the mental health organizations, with a 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 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.

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

In the regressions that follow, 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.

Figure 1 shows scatterplots of expenditures versus beds and additions for the sample of 875 facilities (omitting one extremely large facility described below). In the first row, points are highlighted whose OLS leverage is greater than $2p/n = 0.007$. The second row shows bubbleplots with the relative size of the bubbles proportional to the weight of each case. High SW leverage points are highlighted using the same cutoff of 0.007. The distributions of the predictors are quite skewed as noted in Table 1. There is also one very large facility that is not shown in Figure 1 because it distorts the scale of the plot. That facility (denoted as observation 818 here) has (expenditures in 1,000's; beds; additions) = ($519,863.3; 2,405; 79,808) and has a survey weight of 2.22. (Observation 818 was one of the cases noted earlier that was a certainty in the initial sample but received a nonresponse adjustment, and, thus, had a final weight larger than 1.) Because its data values are far out of line with those of the other organizations, this point has the potential to affect estimates.

Table 2 reports the twenty observations with the largest SW leverages. The values of the leverages range from 0.022 to 0.389, substantially greater than the level of the rough rule of thumb 0.007. This table also shows, for these twenty cases, the OLS unweighted leverages, the ratio of individual sample weight to average sample weight and the relative absolute distance between individual X values and their weighted means. We note that unit 818 has the highest weighted and unweighted leverages, mainly resulting from its extremely large number of beds and number of additions. Since this case has a less-than-average sample weight, the OLS leverage is even larger than the weighted one. There are other similar cases such as units 271, 179, 820, 157, 163, 156, and 154, which are associated with either extreme number of beds, or extreme number of additions, or both – but have small weights. Another type of outlier results from extreme sample weights, even if the values of their auxiliary variables are not very distinct from others. Units 672, 613, 711, 801, and 611 all have sample weights more than 15 times the average weight. Their weighted leverages are identified as large, whereas the unweighted leverages are not. There is also a noticeable gap between the weighted leverages for case 331 ($h_{ii} = 0.075$) and for case 271 ($h_{ii} = 0.046$).

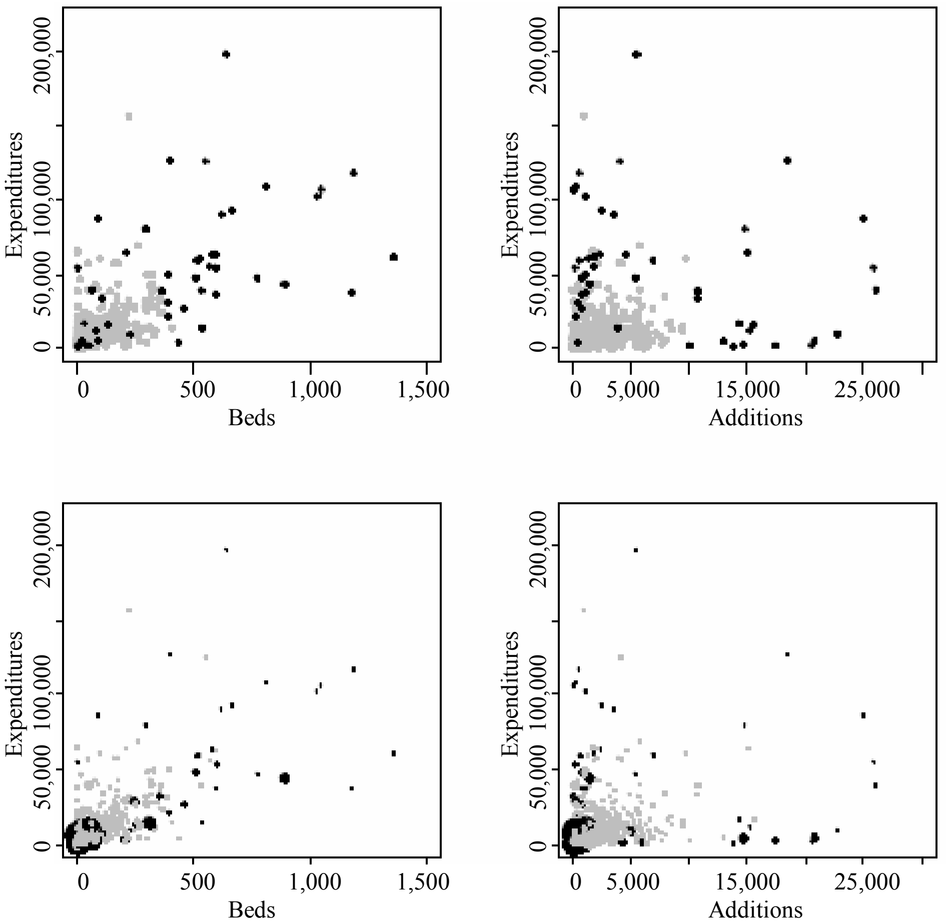

Figure 1 Scatterplots of expenditures versus beds and additions. High leverage points based on OLS (SW) are highlighted in top (bottom) row

Table 2
Observations with 20 largest survey weighted leverages

			Weights	**Beds**	**Additions**
Obs ID	**OLS h_{ii}**	**Weighted h_{ii}**	$w_i / \bar{w}$	$\lvert x_{1i} - \bar{x}_1^W \rvert / \bar{x}_1^W$	$\lvert x_{2i} - \bar{x}_2^W \rvert / \bar{x}_2^W$
818	0.513	0.389	0.3	49.3	64.7
189	0.037	0.245	3.4	17.7	0.3
346	0.035	0.157	2.2	0.6	16.1
366	0.017	0.105	3.0	0.7	11.1
331	0.024	0.075	1.5	0.1	13.4
271	0.068	0.046	0.4	23.7	0.0
830	0.004	0.045	5.8	5.4	0.1
628	0.056	0.045	0.4	1.0	20.3
179	0.089	0.038	0.2	27.4	0.5
672	0.002	0.034	24.2	1.0	0.8
820	0.048	0.034	0.3	0.8	19.6
207	0.012	0.030	1.3	9.5	0.3
157	0.069	0.030	0.2	23.8	0.5
163	0.017	0.027	0.8	11.4	0.8
613	0.002	0.026	18.5	1.0	0.7
711	0.002	0.024	16.8	1.0	0.9
801	0.002	0.024	17.5	0.6	0.9
156	0.055	0.023	0.2	20.9	0.9
611	0.002	0.023	15.9	1.0	0.8
154	0.051	0.022	0.2	20.5	0.1
			$\bar{w} = 6.57$	$\bar{x}_1^W = 47.83$	$\bar{x}_2^W = 1{,}214.13$

Note: observation ID is the line number of an observation in the sample.

Sizes of the sample weights can make analysts reach different conclusions when they use weighted or unweighted leverages to identify potentially influential observations. Figure 2 shows a scatterplot of weighted leverages versus unweighted ones. The two reference lines were drawn at values of 0.007. Observation 818 is omitted since it would again distort the scale of the graph. Clearly, the high leverage points identified by the SW method only, located in area A, have significantly larger weights than the points in area B, which are identified by the OLS method only.

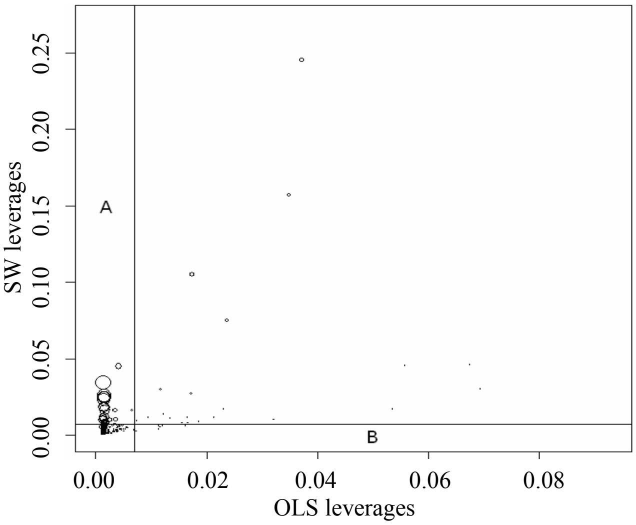

Figure 2 Plot of survey weighted leverages versus OLS unweighted leverages

Given that some potentially influential cases have been identified, the next step is to see what effect they have on parameter estimates. Table 3 shows the OLS and SW parameter estimates using all cases. Table 4 lists the OLS and SW estimates (i) omitting high leverage cases and (ii) omitting observation 818. High leverage points are those with $h_{ii} > 0.007$. However, note that different sets of points are high leverage in OLS and SW regressions. The standard errors are estimated via the usual OLS formula and the sandwich estimator (Binder 1983) for the SW estimates.

Comparing Tables 3 and 4, we see that the OLS estimates change substantially after high leverage points are deleted (section (i) of Table 4). The OLS intercept, which is significant in both tables, jumps from negative to positive. The OLS slope for beds drops by about 26% (94.16 to 69.27) when the high leverage points are dropped. The decrease is about 59% for the slope for additions. The SW estimates for beds and additions are also sensitive to the high leverage points with the slopes decreasing by 7% and 46% respectively. In all cases, the slopes are significant so that the qualitative conclusion that expenditures is related to beds and additions holds with or without the high leverage points. However, predicted values will be quite different before and after omitting these points.

The standard errors (SE's) also decrease substantially when the high leverage points are omitted. For example, the SW standard error for beds drops from 13.14 to 6.75 (a 49% reduction); the SE for additions drops from 0.76 to 0.21 (a 72% reduction). This is due to some points with extreme weights being removed in the SW regression. In contrast, the SE's for the OLS estimates actually increase when the OLS high leverage points are omitted because the sample variance of the *x*'s decreases. This is another illustration of the considerable differences that can occur when applying the same type of diagnostic to OLS and SW regressions.

Table 3
OLS and SW parameter estimates of SMHO regression using all 875 sample cases

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

Table 4
OLS and SW parameter estimates after from SMHO regression

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
(i) Deleting observations with leverages greater than 0.007						
Intercept	2,987.55	490.54	6.09	1,993.86	353.71	5.64
# of Beds	69.27	4.35	15.94	75.82	6.75	11.23
# of Additions	0.95	0.20	4.71	1.00	0.21	4.73
(ii) Deleting observation 818						
Intercept	1,979.51	537.93	3.68	2,281.17	460.35	4.96
# of Beds	81.80	2.92	27.98	68.69	8.04	8.54
# of Additions	1.19	0.14	8.41	0.79	0.29	2.75

Because point 818 is so obviously extreme, we also fitted the regression after dropping only that observation. The results are shown in section (ii) of Table 4. Omitting that single point causes noticeable changes in both OLS and SW parameter estimates. This also illustrates that a single point can affect the standard errors for estimated slopes in a survey-weighted regression, as is also the case in OLS. Observation 818 has a large residual (see Figure 3); omitting it results in the SE for Beds dropping from 13.14 in Table 3 to 8.04 in Table 4. Note that if unit 818 had a large weight, then its residual would likely be smaller since it would have more affect on the fit. If so, the SE could actually be smaller when unit 818 is included.

Another point to be gleaned from Tables 3 and 4 is that the OLS and SW estimates are much closer to each other after the high leverage points are dropped than they are before. As shown in Table 5, the OLS estimates are 16 and 26% larger than the SW estimates with all points but are 9 and 5% less than SW after dropping points.

Table 5
Ratios of OLS and SW parameter estimates before and after deleting observations with leverages greater than 0.007 from SMHO regression

	Ratio of OLS to SW estimates	
	With all points	Dropping high leverage points
Beds	1.16	0.91
Additions	1.26	0.95

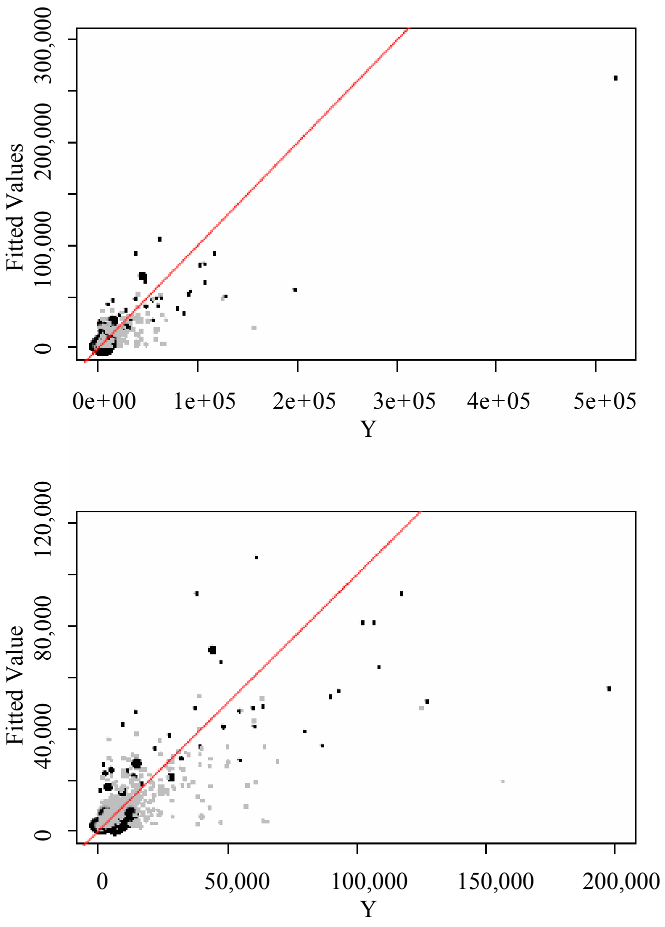

Figure 3 Plot of fitted values versus Y values. Reference line is drawn at $Y = \hat{Y}$. The upper panel includes all points. The lower panel omits the extreme observation 818. High leverage points based on SW are solid, dark circles in each panel

Leverages are usually combined with residuals to determine which points are influential in fitting the regression model because residuals can be used to detect discrepant Y values. A scatterplot of fitted values from the SW regression versus the Y values is shown in Figure 3. The high leverage points are labeled as dark solid circles. The vertical distances from the points to the 45 degree line imply the sizes of the residuals. The upper panel includes all 875 sample points; the lower panel omits observation 818 to provide better resolution for the remaining points. Note that some observations have high leverages and small residuals, while others have low leverages and large residuals. The influence of these points on the regression can be further investigated using various tools that we will not cover here. For example, Cook's distance, implicitly involving the leverage and residual, is designed to measure the effect of deleting a single observation on the overall parameter estimates. The adaptation of some basic OLS diagnostic statistics to survey data, such as DFBETAS and DFFITS, has been discussed under a single stage sampling design in Li and Valliant (2006).

6. Conclusion

Leverages and residuals are essential components of diagnostic statistics intended to identify substantial influence of a single observation or a group of observations on a fitted linear model. Survey data sets can contain influential observations whether one argues that the sample design is ignorable and ordinary least squares can be used, or that the design must be accounted for and survey weights used. The points that are influential in the two cases are not necessarily the same, as illustrated here.

Once high leverage points are identified, an important question is how to deal with them for inference. Two options are to down-weight them or drop them from model-fitting entirely. Down-weighting seems unsatisfactory in general since a point can have a high leverage not because of a large weight but rather due to having one or more unusual *X*'s. Down-weighting may be sensible from a model-based point-of-view, assuming the model itself is correctly specified. However, the design-based idea of estimating a census parameter may then be lost. If a point has a large leverage because of extreme *X*'s, then it may not follow the model at all and should be dropped.

However, using a mechanical procedure that automatically drops many influential observations with high leverages can lead to standard error estimates that are too small, resulting in confidence intervals that cover at less than the nominal rates and in inflated Type I error rates in hypothesis tests (Li 2007). This phenomenon is similar to well-known problems in stepwise regression (Hurvich and

Tsai 1990, Zhang 1992). Thus, a useful research topic appears to be developing inferential procedures for constructing confidence intervals and conducting hypothesis tests that account for the effects of dropping or down-weighting points.

For complex survey data, the hat matrix involves no design features except for sample weights and can be used to identify cases that have atypical weights or predictor values. Other diagnostic statistics, like Cook's D, do contain variance estimates that need to account for complex sample design features such as stratification and clustering. The adaptation and extension of additional diagnostic approaches for survey analysis will be explored in the future.

7. Acknowledgement

This material is based 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 author(s) and do not necessarily reflect the views of the National Science Foundation. We thank the editor and referees for their thoughtful comments, which improved the paper considerably.

Appendix

Inclusion of certainties in standard error estimation

In the empirical study in section 5, we included certainty units in the standard error calculations. The justification for doing this is sketched here. Under the general model (3), the model variance of $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, the estimator used in the empirical study, is $\text{var}_M(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\mathbf{A}^{-1}\sigma^2$ where $\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and $\mathbf{V} = \text{diag}(v_i)_{i\in s}$. The sandwich variance estimator used in the study reported in section 5 is defined as

$$v(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\frac{n}{n-1}\sum_{i\in s}(\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})^T\mathbf{A}^{-1} \qquad (4)$$

where $\mathbf{z}_i = w_i e_i \mathbf{x}_i$ with $e_i = Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ and $\overline{\mathbf{z}} = \sum_{i\in s} w_i e_i \mathbf{x}_i / n$. This estimator is design consistent (see Binder 1983) in single-stage sampling if units are sampled with replacement with probabilities equal to w_i^{-1}, and there are no certainty units. If the sample contains certainties, the formula for $v(\hat{\boldsymbol{\beta}})$ would be modified to estimate the design-based variance: certainties would be excluded from the sums in (4) and $\overline{\mathbf{z}}$, and n would be changed to n_{nc}, the number of non-certainties. In the extreme case of a census, the design-based variance estimator would reduce to zero.

The estimator in (4) is approximately model-unbiased under (3) regardless of whether the sample contains certainties or not. The middle matrix in (4) can be expanded as $\sum_{i\in s}(\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})^T = \sum_{i\in s}\mathbf{z}_i\mathbf{z}_i^T - n\overline{\mathbf{z}}\overline{\mathbf{z}}^T$. Assuming that $e_i \approx Y_i - \mathbf{x}_i^T\boldsymbol{\beta}$, the model expectation under (3) of the first term is $E_M(\sum_{i\in s}\mathbf{z}_i\mathbf{z}_i^T) = \mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\sigma^2$ while $E_M(n\overline{\mathbf{z}}\overline{\mathbf{z}}^T) = n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\sigma^2$. Substituting these expectations gives $E_M[v(\hat{\boldsymbol{\beta}})] = \text{var}_M(\hat{\boldsymbol{\beta}})$, which holds even when some units are certainties. This also shows that $v(\hat{\boldsymbol{\beta}})$ is robust in the sense of properly reflecting the contribution of heteroscedastic variances in (3) to the model-variance of $\hat{\boldsymbol{\beta}}$ even though $\mathbf{V}$ may be unknown and not accounted for in the estimation of $\boldsymbol{\beta}$.

References

Beaumont, J.-F., and Alavi, A. (2004). Robust Generalized Regression Estimation. *Survey Methodology*, 30, 195-208.

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

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D.A., and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters, Chapter 3 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.

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

Chambers, R.L., Dorfman, A.H. and Sverchkov, M.Y. (2003). Nonparametric regression with complex survey data, Chapter 11 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.

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

Choudhry, G. (2000). The 1998 Survey of Mental Health Organizations Survey Design. Westat technical report prepared for Center for Mental Health Services, Substance Abuse and Mental Health Services Administration (SAMHSA), available by request to SAMHSA.

Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15-18.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.

Gwet, J., and Rivest, L. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

Hoaglin, D.C., and Welsch, R.E. (1978). The hat matrix in regression and ANOVA (Corr: 78V32 p146). *The American Statistician*, 32, 17-22.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.

Hurvich, C.M., and Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214-217.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

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

Li, J., and Valliant, R. (2006). Influence analysis in linear regression with sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3330-3337.

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

Moreno-Rebollo, J.L., Muñoz-Reyes, A. and Muñoz-Pichardo, J. (1999). Influence diagnostic in survey sampling: Conditional bias. *Biometrika*, 86, 923-928.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.

Potter, F.J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.

Potter, F.J. (1993). The effect of weight trimming on nonlinear survey estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758-763.

Skinner, C.J. (2003). Introduction to Part B, Chapter 6 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.

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

Smith, T.M.F. (1987). Influential observations in survey sampling. *Journal of Applied Statistics*, 14 , 143-152.

Smith, T.M.F. (1989). Introduction to Part B, Chapter 6 in *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley & Sons, Inc.

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

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

Welsh, A.H., and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society*, Series B, Methodological, 60, 413-428.

Zaslavsky, A.M., Schenker, N. and Belin, T.R. (2001). Downweighting influential clusters in surveys: Application to the 1990 post enumeration Survey. *Journal of the American Statistical Association*, 96 , 858-869.

Zhang, P. (1992). Influence after variable selection in linear regression models. *Biometrika*, 79, 741-746.