Influence Analysis in Linear Regression with Sampling Weights

Jianzhu Li[1], Richard Valliant[2]
Joint Program of Survey Methodology, University of Maryland, College Park[1]
Survey Research Center, University of Michigan, Ann Arbor[2]

Abstract

This paper is concerned with the identification of influential observations when we analyze survey data using a linear regression estimator involving survey weights. Based on conventional OLS diagnostic approaches, adapted statistics are proposed and justified to deal with the sampling weights in survey data. Using a sample from NHANES data, a comparison will be made between the diagnostics with and without sampling weights for some types of diagnostic statistics.

Keywords: Regression, Influential observations, Sampling weights, Sandwich Variance Estimator, Linearization Variance Estimator, Leverage, Residual, DFBETAS, DFFITS

1. Introduction

Several decades have passed since linear regression analysis became a widely employed statistical methodology that utilizes the relation between quantitative and qualitative variables to make predictions and inferences. Discussion of diagnostics for linear regression models are often indispensable chapters or sections in most of the statistical textbooks on linear models. Although techniques for regression diagnostics have been developed theoretically and methodologically for conventional linear regression models, diagnostics have not been extensively studied in survey sampling. This paper is concerned with the identification of influential observations when we analyze survey data using a linear regression estimator involving survey weights. A comparison will be made between the diagnostics with and without sampling weights.

Examples from real surveys show that there is a need for influence diagnostics since a small number of the sampled units with extreme values could play a crucial role in the estimation of statistics and their variances. Chambers (1986) characterized outliers in survey data into nonrepresentative and representative. The premise in this research is that an analyst will be looking for a model that fits reasonably well for the bulk of the population. 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, or their estimated standard errors, or both. Cook and Weisberg (1982) propose that the basic idea in influence analysis is to monitor how small perturbations change the outcome of the analysis when they are introduced in the data. 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 and DFFITS, etc. These statistics need to be adapted for application to randomization inference for sample surveys.

The influence of observations on regression estimation under the survey setting may come from at least three sources: outlying **Y** values, **X** values, and sampling weights $\mathbf{W}=(w_1,...,w_n)^T$. Atypical or extreme values of any of these or a combination of these can affect both parameter estimates and their estimated standard errors. Unlike conventional model-based influence diagnostics which have been available in standard software for ordinary least squares, diagnostics for regression using complex survey data need to pay attention to the following: First, as a source of influence, survey weights, which are computed with the primary goal of estimating finite population statistics, need to be incorporated into the construction of influence measurement. Second, the model assumptions which provide the basis of justification for conventional influence diagnostics are partially violated or completely ignored in the context of randomization inference. Third, given the large sample size in many surveys it would be important to set up some criteria to single out the influential units, instead of reporting diagnostics for all units in the sample. Belsley, Kuh and Welsch (1980) recommended choosing reasonable cutoffs by judgment and intuition, combining empirical and theoretical arguments.

2. Regressions for Complex Survey Data

Parameter estimates in linear regression using survey data are derived from the Pseudo Maximum Likelihood (PML) approach (Skinner, Holt, and Smith, 1989). Suppose that the underlying structural model is a fixed-effects linear model:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \text{ind}\, N(0, \sigma^2) \qquad (1)$$

where ε_i is independently normally distributed with mean 0 and variance σ^2. The model-based likelihood for $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}) = \prod_{i \in s} f(Y_i; \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2),$$

where s is the set of sample units and $f(Y_i; \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)$ is the normal density with mean $\mathbf{x}_i^T \boldsymbol{\beta}$ and variance σ^2. The PML estimate of $\boldsymbol{\beta}$ is the solution to the set of estimation equations $\sum_{i \in s} w_i \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, where w_i is the survey weight for unit *i*. Survey weights, which in probability samples are usually inversely proportional to inclusion probabilities, are used in PML to account for an informative design in which sample distribution of the $\mathbf{Y}$'s is likely to differ from that of the finite population. The estimation equations based on the normal probability density function can be simplified as $\mathbf{X}^T \mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$ and solved explicitly as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$.

The above regression estimator $\hat{\boldsymbol{\beta}}$ is approximately design unbiased for the finite population parameter $\mathbf{B} = (\mathbf{X}_N^T \mathbf{X}_N)^{-1} \mathbf{X}_N^T \mathbf{Y}_N$, where $\mathbf{Y}_N = (Y_1, ..., Y_N)^T$, and $\mathbf{X}_N^T = (\mathbf{x}_1, ..., \mathbf{x}_N)$. It is also unbiased for the superpopulation slope $\boldsymbol{\beta}$ in model (1), regardless of whether the variance is specified correctly or not. The finite population parameter $\mathbf{B}$ should be close to $\boldsymbol{\beta}$ for a good model, and therefore a design-based estimate of $\mathbf{B}$ should also estimate $\boldsymbol{\beta}$. This estimator will be referred as Survey Weighted estimator (SW) 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 really follow the model. In that case, an OLS estimator 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 because of its dependency on variables of interest.

Using sampling weights in 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. Although both $\hat{\boldsymbol{\beta}}$ and the OLS estimator $\mathbf{b}$ are model-biased estimators for $\boldsymbol{\beta}$ when necessary covariates are omitted, the model bias of $\hat{\boldsymbol{\beta}}$ diminishes while the sample size increases, whereas $\mathbf{b}$ is only asymptotically unbiased if the selection probabilities are not related to the variables that are left out of the model. The advantage of using the weighted estimates 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). In this study, we assume that analysts will use survey weights to estimate regression models. The diagnostics to be developed account for the effects of these weights.

3. Variance Estimation

As in OLS influence diagnostics, some statistics are formulated using variance estimates of $\hat{\boldsymbol{\beta}}$ and cutoff points are developed in terms of some distributions. In this study we consider the variance estimate of $\hat{\boldsymbol{\beta}}$ under a design of single stage sampling with replacement and with varying probabilities. Assuming $\max(w_i n / N) = O(1)$, where n and N are sample size and population size, respectively, we have

1) $\mathbf{A} = \mathbf{X}^T \mathbf{W} \mathbf{X} = O(N)$, elementwise.
2) $\mathbf{C} = \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} = O(n^{-1})$, elementwise.
3) $\mathbf{H} = \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} = O(n^{-1})$, elementwise.

If the working model is (1), treating the finite population as a sample of size *N* for that model, we estimate the model error variance σ^2 using

$\sigma_U^2 = \sum_{i \in U} \frac{e_{iU}^2}{N-p}$, where $e_{iU} = y_i - \mathbf{x}_i^T \mathbf{B}$. σ_U^2 is an model unbiased estimate of σ^2 . According to the pseudo maximum likelihood approach, we can obtain the design-based estimate of σ_U^2 from a sample of size n using a π-estimator

$$\hat{\sigma}^2 = \frac{1}{\hat{N}} \sum_{i \in s} w_i e_i^2 \qquad (2)$$

where e_i is the sample residual defined as $e_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$. As sketched in the Appendix, $\hat{\sigma}^2$ is an approximately design unbiased estimator for σ_U^2 and, if the working model is correctly specified, is also estimating σ^2 .

Suppose that an analyst uses the survey-weighted estimator $\hat{\boldsymbol{\beta}}$, which can be rewritten as a weighted sum of the $\mathbf{Y}$ values, $\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{A}^{-1} \mathbf{x}_i w_i Y_i$. Its unknown model variance is

$$\text{var}_M\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \mathbf{A}^{-1} \left(\sum_{i=1}^{n} w_i^2 \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{A}^{-1},$$

which can be estimated as

$$v_M\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \mathbf{A}^{-1} \left(\sum_{i=1}^{n} w_i^2 \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{A}^{-1}. \qquad (3)$$

If the variance structure of model (1) is misspecified, instead let us consider a more general model in which the Y_i 's are independent but whose variance differs among the units:

$$Y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i \sim ind(0, \psi_i), \qquad (4)$$

where ψ_i is an unknown variance parameter. The model variance of $\hat{\boldsymbol{\beta}}$ is

$$\text{var}_M\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{A}^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i w_i \psi_i w_i \mathbf{x}_i^T \right) \mathbf{A}^{-1}. \qquad (5)$$

Under model (4), the squared residual has expectation

$$E_M\left(e_i^2\right) = \psi_i \left(1 - h_i\right)^2 + \sum\nolimits_{j \neq i} h_{ij}^2 \psi_j,$$

with h_{ij} being the (ij)th element of the hat matrix $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}$. Under certain regularity conditions, asymptotically $E_M\left(e_i^2\right) \approx \psi_i$ and therefore e_i^2 is an approximately model-unbiased estimator of ψ_i (Valliant, Dorfman, and Royall 2000). By replacing the unknown variance elements ψ_i in (5) by the squares of the corresponding residuals e_i^2 based on the regression fit, the sandwich estimator of the unknown model variance is

$$v_W\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{A}^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i w_i e_i^2 w_i \mathbf{x}_i^T \right) \mathbf{A}^{-1}. \qquad (6)$$

Using $\mathbf{C} = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{W} = \left(c_{jk}\right)_{p \times n}$, we have $v_W(\hat{\boldsymbol{\beta}}_j) = \sum_{k=1}^{n} c_{jk}^2 e_k^2$. This estimator is model robust against deviations from the constant variance structure as in model (1). It is also design consistent under a single-stage, unstratified and unclustered design where units are selected with probabilities, $\pi_i = 1/w_i$, with replacement.

Another useful variance estimator is the design-based linearization variance estimator. The linear approximation of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} - \mathbf{B} \square\ \mathbf{A}_N^{-1} \sum\nolimits_{i \in s} \mathbf{x}_i^T w_i (Y_i - \mathbf{x}_i^T \mathbf{B}) = \sum\nolimits_{i \in s} \mathbf{z}_i \qquad (7)$$

where $\mathbf{A}_N = \mathbf{X}_N^T \mathbf{X}_N$, and $\mathbf{z}_i = \mathbf{A}_N^{-1} \mathbf{x}_i w_i (Y_i - \mathbf{x}_i^T \mathbf{B})$ (Fuller, 2002). If the design is approximated by single stage with-replacement sampling, the linear substitute approach can be used to obtain the design consistent variance estimator

$$v_L\left(\hat{\boldsymbol{\beta}}\right) = \frac{n}{n-1} \sum_{i=1}^{n} (\mathbf{z}_i^* - \bar{\mathbf{z}}^*)(\mathbf{z}_i^* - \bar{\mathbf{z}}^*)^T ,$$
$$= \frac{n}{n-1} \sum\nolimits_{i=1}^{n} \mathbf{z}_i^* \mathbf{z}_i^{*T}$$

where $\mathbf{z}_i^* = \mathbf{A}^{-1} \mathbf{x}_i w_i e_i$, $e_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, and $\bar{\mathbf{z}}^* = \sum_s \mathbf{z}_i^* / n = \mathbf{0}$ (e.g. see SUDAAN v.8 manual). This estimator also has robust model-based interpretations under certain types of misspecification of model variance parameter. v_L and v_W are approximately the same when the sample size is large enough that $\frac{n}{n-1} \approx 1$.

4. Adaptations of Traditional Diagnostics to Regression on Survey Data

4.1 Residuals and Leverages

When survey weights are used in the regression, the predicted values become $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ and the residuals are $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. The survey weighted hat matrix $\mathbf{H}$ has the following properties:

1) $0 \le h_i \le 1$;

2) $\sum_{i=1}^{n} h_i = p$,

where p is the number of columns in $\mathbf{X}$ matrix. (for proofs see Valliant et. al., 2000). Leverages are defined as the diagonal elements of the hat matrix, which are the weights of observation Y_i in determining the fitted value $\hat{Y}_i$. A large leverage may be caused by outlying $\mathbf{X}$ values, an outlying weight, or both. However, a large residual mainly results from an outlying Y_i.

Leverages can be decomposed into components that separate the effect of the weight and the $\mathbf{X}$ values for a unit. Assuming we have a model with intercept, let

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

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

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

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 $\mathbf{A}_1 = \mathbf{X}_1^T\mathbf{W}\mathbf{X}_1$ is a $(p-1)\times(p-1)$ matrix. Using the inverse of partitioned matrix, we have

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

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

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

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

$$= \left(\mathbf{1}\ \mathbf{X}_1\right)\left\{\begin{pmatrix} \frac{1}{\hat{N}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\bar{\mathbf{x}}_W^T \\ \mathbf{I} \end{pmatrix}\mathbf{S}^{-1}\left(-\bar{\mathbf{x}}_W\ \mathbf{I}\right)\right\}\begin{pmatrix} \mathbf{1}^T \\ \mathbf{X}_1^T \end{pmatrix}\mathbf{W}$$

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

Then the leverage of ith observation, or the ith diagonal element of $\mathbf{H}$, is

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

$\left(\mathbf{x}_i - \bar{\mathbf{x}}_W\right)^T\mathbf{S}^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}_W\right)$ is an ellipsoid centered at $\bar{\mathbf{x}}_W$ (e.g., see Weisberg 1985). 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 of the $\mathbf{X}$'s, $\bar{\mathbf{x}}_W$.

Usually it is helpful to standardize the residuals for residual analysis. In the OLS case, a residual is scaled either by $\sqrt{\text{MSE}}$ or by its estimated standard error to obtain a semistudentized or studentized residual.

Assuming single stage sampling, under model (1), the residual for unit i is $e_i = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ and its model variance is $E_M\left(e_i^2\right) = \sigma^2\left[\left(1-h_i\right)^2 + \sum_{j\ne i} h_{ij}^2\right]$. Since $h_{ij} = O(n^{-1})$, and $E_M\left(e_i^2\right) \square\ \sigma^2$, we can standardize the residual for unit i by $\hat{\sigma}$ estimated from (2) and compare it with a standard normal random variable.

It is not feasible to define the distribution of residuals from the design-based point of view. However, plots of residuals are helpful in highlighting data points suspected of unduly affecting the fit of regression. The added variable plot, also known as partial regression leverage plot, provides a method of assessing the impact of individual observations on the estimate of a single parameter $\hat{\beta}_k$ in a multiple regression model. Korn and Graubard (1999) illustrated the use of these plots with survey data. Let $\mathbf{X}(-k)$ be $n\times(p-1)$ matrix formed from the data matrix, $\mathbf{X}$, by removing its kth column, $\mathbf{x}_k$. Further let $\mathbf{u}_k$ and $\mathbf{v}_k$ be the residuals that result from regressing $\mathbf{Y}$ and $\mathbf{x}_k$ on $\mathbf{X}(-k)$ using survey weights. The kth regression coefficient of the multiple regression model, $\hat{\beta}_k$, is the same as the slope coefficient of the weighted regression of $\mathbf{u}_k$ on $\mathbf{v}_k$. The added variable plot is defined as a scatter plot of $\mathbf{u}_k$ against $\mathbf{v}_k$ along with their simple linear regression line. For survey data it is presented as a bubble plot with each bubble representing an observation and its area proportional to the sample weight. Although the plot is not able to precisely measure how severely an observation is different from others, when it is used as an extra tool to the adapted methodologies, it can directly tell us why some points are identified as outlying and toward which direction those points pull the weighted regression line.

4.2 DFBETAS

DFBETA is the change in parameter estimate after deleting the *i*th observation. It becomes

$$DFBETA_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{\mathbf{A}^{-1}\mathbf{x}_i^T e_i w_i}{1-h_i},$$

$$\text{and } DFBETA_{ij} = c_{ji} e_i / (1-h_i),$$

when sampling weights $\mathbf{W}$ are taken into consideration (Valliant, et al 2000). It is different from the one in the OLS case in both numerator and denominator because sample weights are involved in the leverages and residuals. To create a survey weighted version of DFBETAS, we need to divide DFBETA by an estimate of the standard error of $\hat{\boldsymbol{\beta}}$ that accounts for sample weights. Under model (1), we propose a specification of DFBETAS statistic as follows:

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

$$= \frac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c_{jk}^2}} \cdot \frac{e_i}{\hat{\sigma}} . \frac{1}{\sqrt{1-h_i}} .$$

Knowing the order conditions $c_{jk} = O(n^{-1})$ and $h_i = O(n^{-1})$ and assuming that $\frac{e_i}{\hat{\sigma}}$ is approximately $N(0,1)$ in large samples, we rewrite the DFBETAS statistic as the approximate product of two terms, $DFBETAS_{ij} \approx O(n^{-1/2}) \cdot N(0,1)$. An observation i may be identified as influential on the estimation of $\hat{\beta}_j$ if $\left| DFBETAS_{ij} \right| \geq \frac{2}{\sqrt{n}}$. This is the same cutoff suggested by Belsley et al (1980) for OLS. Moreover, the model robust sandwich estimator $v_W(\hat{\beta}_j)$ or the linearization variance estimator $v_L(\hat{\beta}_j)$ can be used to replace $v_M(\hat{\beta}_j)$ in case the underlying model deviates from the working model.

4.3 DFFITS

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

$$DFFIT_i = \hat{Y}_i - \hat{Y}_i(i) = \mathbf{x}_i^T \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} \right) = \frac{h_i e_i}{1-h_i} .$$

The model variance of $\hat{Y}_i$ is

$$Var_M\left(\hat{y}_i\right) = \sigma^2 \left(\mathbf{HH}^T \right)_{ii} = \sigma^2 \sum_k h_{ik}^2 ,$$

which is estimated by $v_M\left(\hat{y}_i\right) = \hat{\sigma}^2 \sum_k h_{ik}^2$. In OLS, $\sum_k h_{ik}^2 = h_i$ because $\mathbf{HH}^T = \mathbf{H}$ 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 $v_M\left(\hat{y}_i\right)$ and rearranged as follows:

$$DFFITS_i = \frac{h_i e_i / (1-h_i)}{\sqrt{v_M\left(\hat{y}_i\right)}} = \frac{h_i e_i / (1-h_i)}{\sqrt{\hat{\sigma}^2 \sum_k h_{ik}^2}}$$

$$= \sqrt{\frac{h_i}{1-h_i}} \frac{e_i}{\hat{\sigma}} \frac{1}{\sqrt{\sum_k h_{ik}^2 / h_i}} \frac{1}{\sqrt{1-h_i}}$$

$$\cong \sqrt{\frac{p}{n}} \cdot N(0,1) \cdot O(1),$$

where we approximate h_i by its mean p/n and use $h_i / \left(1-h_i\right) \approx p / \left(n-p\right) \approx p/n$. Hence the cutoff value is $2\sqrt{\frac{p}{n}}$ if we use DFFITS to determine the influential observations. Note that this is also the same cutoff suggested for OLS by Belsley et al (1980).

5. Case Study

The adapted diagnostic techniques will be applied and justified based on a linear regression analysis of data from the National Health and Nutrition Examination Survey (NHANES). This survey is a rich source of quantitative and qualitative variables and is quite important for analysis of health conditions in the U.S. There are several of these data sets publicly available. From NHANES 1999-2002 we draw a sample of size

500 among women aged 20 to 39. In order to have a large variation in sample weights, we keep the 100 observations with the largest weights and the 100 observations with the smallest weights, and select the remainder 300 randomly. The survey weights in this subset range from 698.39 to 103831.17. This set was selected to illustrate the use of the diagnostics and does not represent any particular estimation domain. The influence analysis is conducted on the regression of systolic blood pressure on the logarithm of blood lead levels, ages, and body mass index. The same regression analysis has been done by Korn and Graubard (1999) using a different sample. The stratified clustering design of the survey is ignored in our diagnostics.

5.1 Parameter Estimation

Table 1 shows the difference of parameter estimates between the OLS estimation and the Survey Weighted estimation. The coefficients of age and BMI have slightly discrepancies between the two methods whereas the coefficient of blood lead differs dramatically. The effect of survey weights on coefficient estimation signals that survey weights could play a crucial role in influence analysis on this regression.

Table 1. Coefficient Estimates and Their Standard Errors for the Ordinary Least Squares Estimation and the Survey Weighted Estimation.

Independent	OLS Estimation		SW Estimation	
Variables	Coefficient	SE	Coefficient	SE
Intercept	91.803	2.853	89.575	3.319
Age(years)	0.060	0.075	0.098	0.095
BMI	0.536	0.068	0.603	0.101
Lead (Log)	1.367	0.714	2.262	1.053

5.2 Diagnostics by Leverages and Residuals

Figure 1 is a scatterplot of leverages calculated using two methods: the OLS formulation and the survey weighted formulation. Influential leverages, greater than twice their mean, are identified to be associated with the points beyond the two reference lines. The 54 influential observations identified by SW but not by OLS diagnostics are associated with large sample weights ranging from 31655.76 to 103831.17, which are represented by the size of the bubbles; whereas the 23 influential observations identified by OLS only have small weights, ranging from 725.10 to 17987.69. The bubbles in the upper right square, with moderate sizes, stand for the points identified by both methods.

The points in Figure 2 show the residuals scaled by the estimated standard error $\hat{\sigma}$ of model (1), where $\hat{\sigma}$ is estimated by both the OLS estimator and the survey weighted formula (2). With a few exceptions, the weighted and unweighted diagnostics identify almost the same extreme residuals. The residual analysis mainly filters out the observations with outlying Y values, but not outlying weights.

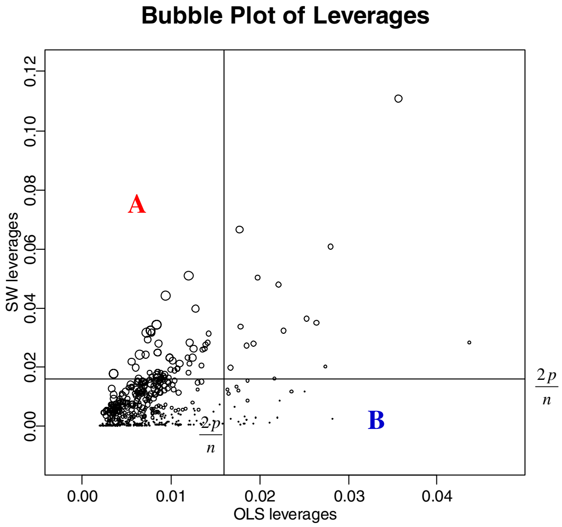

Figure 1. Leverage Plot. Area A includes points identified as influential by the SW diagnostic only. Area B includes points identified by the OLS diagnostic only.

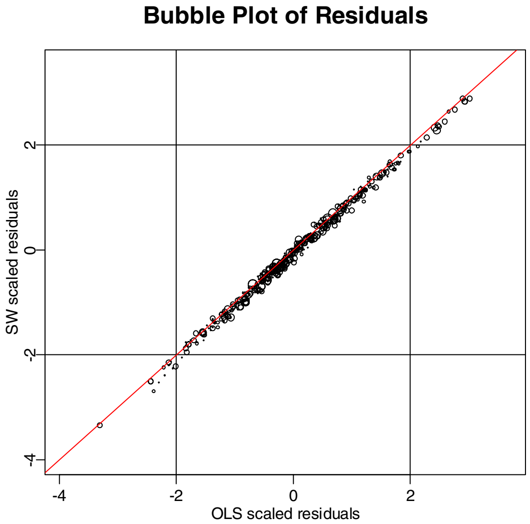

Figure 2. Residual Plot. Scaled residuals with absolute values exceeding 2 indicate the existence of outlying observations.

5.3 Diagnostics by DFBETAS and DFFITS

The diagnostic results of DFBETAS statistics are presented in Figure 3. It conveys the same messages as the leverage diagnostics in Figure 1. Using the SW formula of DFBETAS, we are more likely to single out the points associated with large sampling weights. It is clearly shown in the graph that points identified by the OLS method only have small weights symbolized by the bubbles of small sizes.

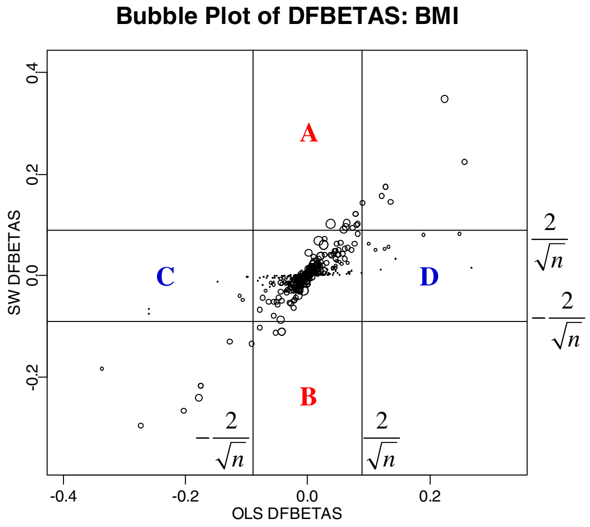

Figure 3. DFBETAS Plot. Areas A and B include points identified only by the SW diagnostics whereas areas C and D include points identified by the OLS diagnostics only.

Another way to show how the deletion of an observation affects the coefficient estimation is to draw an added variable plot, as we introduced in Section 4.1. Figure 4.a and 4.b display two added variables plots of BMI for the OLS regression and the SW regression, respectively. The points in red indicate the influential observations identified by using DFBETAS statistics. In Figure 4.a, the identified influential points are scattered around the corners where they deviate further from the middle of the regression line than the unidentified points. However, in Figure 4.b, the red dots are not necessarily the furthest away from the center of the regression line if they are associated with very large sampling weights. Even some points that stray greatly from the rest are not identified because their weights are too small.

Basically the DFFITS diagnostics reach the same conclusion as DFBETAS unless they identified fewer influential points because DFFITS summarizes the effect of deleting a specific unit on the overall parameter estimation. There are 19 influential observations identified by the SW in Figure 5 but not by the OLS diagnostics, with their weights ranging from 44843.48 to 103831.17. There are 15 influential observations identified by the OLS diagnostics only. Their weights are relatively small, ranging from 833.35 to 31722.48.

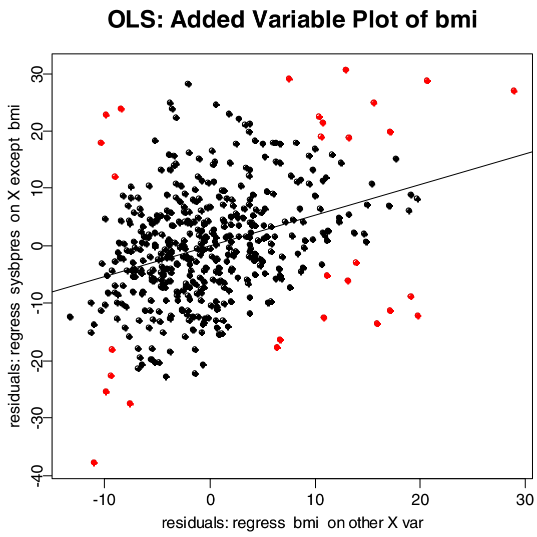

Figure 4.a. Added Variable Plot of BMI using the OLS regression.

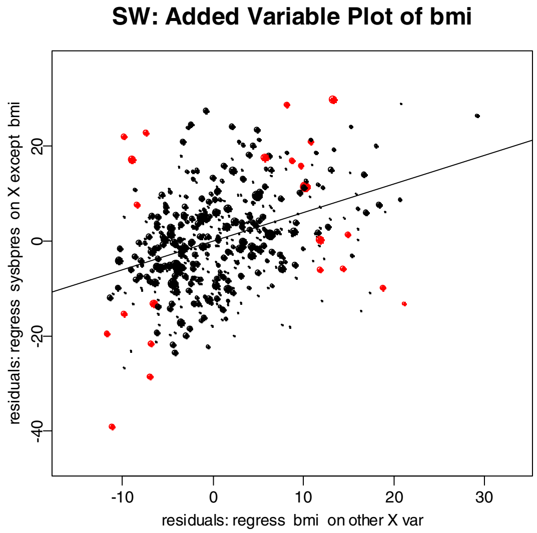

Figure 4.b. Added Variable Plot of BMI using the SW regression.

6. Conclusion

The conventional OLS influence diagnostics are adapted to be used for survey data. The cutoff values for adapted statistics are determined and justified in terms of model distributions and the order of magnitude of survey weights and other sample quantities. Based on the comparison of the OLS and the SW influence analysis on a NHANES sample, we conclude that the SW diagnostics, including leverages, DFBETAS, and DFFITS, identify different points than the OLS diagnostics as being influential. This is because in survey weighted regressions, points can be influential due to outlying sample weights besides extreme Y and $\mathbf{X}$ values.

In this study we only consider the single-stage with replacement sampling designs. In subsequent work we will generalize DFBETAS and DFFITS for a complex sampling design accounting for clustering and stratification. Also, we plan to adapt other statistics used in the conventional OLS diagnostics such as Cook's Distance and COVRATIO, and the expansion of the single-case deletion to the identification of influential groups.

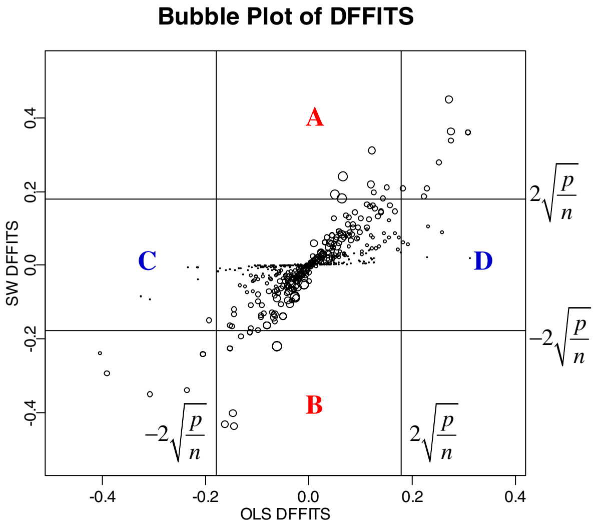

Figure 5. DFFITS Plot. Areas A and B include points identified only by the SW diagnostics whereas areas C and D include points identified by the OLS diagnostics only.

Appendix

Assume that $\hat{\boldsymbol{\beta}} = \mathbf{B} + O_p\left(1/\sqrt{n}\right)$. It follows that

$$\hat{\sigma}^2 = \frac{1}{\hat{N}}\sum_{i\in s} w_i e_i^2 = \frac{1}{\hat{N}}\sum_{i\in s} w_i\left(y_i - \mathbf{x}_i^T\mathbf{B}\right)^2 + O_p\left(1/\sqrt{n}\right),$$

$$E_\pi\left(\hat{\sigma}^2\right) \square \frac{1}{N}\sum_{i\in U}\left(y_i - \mathbf{x}_i^T\mathbf{B}\right)^2 \square \sigma_U^2 \text{ , and}$$

$$E_M\left(\hat{\sigma}^2\right) \square \frac{1}{\hat{N}}\sum_{i\in s} w_i E_M\left(y_i - \mathbf{x}_i^T\mathbf{B}\right)^2 = \sigma^2$$

References

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

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

Cook, R. D., and Weisberg, S. (1982), *Residuals and influence in regression*, Chapman & Hall Ltd (London; New York)

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

Fuller, W. A. (2002), "Regression estimation for survey samples", *Survey Methodology*, Vol. 28, No. 1, 5-23

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

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

Kott, P. S. (1991), "A model-based look at linear regression with survey data", *American Statistician* 45: 107-112

Little, R. J. A. (1991), "Inference with survey weights", *Journal of Official Statistics*, 7: 405-424

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*, ser. A, 148: 268-278

Rubin, D. B. (1985), "The use of propensity scores in applied Bayesian inference", in *Bayesian Statistics* 2, edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith. Amsterdam: North Holland.

Skinner, C. J., Holt, D. and Smith, T. M. F. (eds.) (1989), *Analysis of Complex Surveys*, New York: 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, edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith. Oxford, Eng: Oxford University Press.

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

Weisberg, S. (1985), *Applied linear regression*, New York: Wiley