

## Linear Regression Diagnostics in Cluster Samples

Jianzhu Li

Westat, 1650 Research Boulevard, Rockville MD 20852

### Abstract

This paper extends and adapts the conventional ordinary least squares influence diagnostics to linear regression using complex survey data, following the preliminary study which has been conducted before. Diagnostic statistics such as DFBETAS, DFFITS, and modified Cook's Distance are constructed under complex sampling designs to evaluate the effect on the regression coefficients of deleting a single observation. The forward search method is also adapted to identify influential observations as a group when there is possible masked effect among the outlying observations.

**KEY WORDS:** Influential observations, Sample weights, Forward Search.

### 1. Introduction

Linear regression models and estimators have been widely applied to analyze complex survey data since the Pseudo Maximum Likelihood (PML) approach was outlined and specified (e.g., Binder 1983; Skinner et al. 1989). Using the sample weights in the regression estimator not only allows the analysts to account for the design features which govern the data collection process, but also provides a limited type of robustness to model misspecification (Pfeffermann and Holmes 1985; DuMouchel and Duncan 1983; Kott 1991). The sandwich estimator (Valliant et al. 2000) and the Taylor Series linearization estimator (Fuller 2002) are often employed to obtain both design and model consistent estimated variances of the regression parameters.

Insufficient attention has been given to diagnosing the adequacy of the working models, more specifically, detecting outlying and influential observations for regressions using complex survey data. Previous survey literature has covered locating and trimming extreme sample weights, Many survey literatures discussed locating and trimming extreme sample weights, measuring the effect of outliers on the estimation of descriptive population statistics, and constructing outlier-robust estimation techniques. Li (2007) contains an extensive review of this literature. Elliott (2007) and Korn and Graubard (1999) are two of the few references which introduce techniques for the evaluation of the quality of regression on complex survey data. Li and Valliant (2006) conducted preliminary studies in adapting the traditional OLS diagnostic statistics such as leverages,

residuals, DFBETAS, and DFFITS to survey-weighted (SW) regression estimators assuming the sample was drawn from a single-stage non-stratified with-replacement design. In this paper we will extend our research to more complex sample designs involving stratification and clustering so as to enhance the comprehensiveness and the applicability of the proposed approaches.

### 2. Model Specification and Variance Estimation

To formulate regression diagnostics for clustered survey data, some model structure is needed. Parameters of the model can be estimated as discussed later in this section. We only sketch theoretical results here; details can be found in Li (2007). Suppose that a two-stage sample of units is selected with the first stage sampled clusters or primary sampling units (PSUs),  $n$ , being selected with replacement. Let  $m_i$  be the number of sampled units in the  $i$ th cluster,  $w_{ik}$  be the sample weight of the  $k$ th units in the  $i$ th cluster, and  $N$  the number of PSUs in the population. Suppose there are  $i = 1, \dots, N$  clusters in the population and  $k = 1, \dots, M_i$  units in cluster  $i$ . Suppose that  $\mathbf{x}_{ik}$  is a  $p$ -vector of explanatory variables for unit  $k$  in cluster  $i$ . The linear model is

$$Y_{ik} = \mathbf{x}_{ik}^T \boldsymbol{\beta} + \varepsilon_{ik}$$

$$\text{Cov}_M(\varepsilon_{ik}, \varepsilon_{i'k'}) = \begin{cases} \sigma^2 & i = i', k = k' \\ \sigma^2 \rho & i = i', k \neq k' \\ 0 & i \neq i', k \neq k' \end{cases} \quad (1)$$

This model posits that all units have a common variance and the intraclass correlation,  $\rho$ , is the same for all clusters. Units in different clusters are uncorrelated. In practice,  $\rho$  is usually positive and can be estimated using analysis of variance or related methods.

Using the PML approach, the survey weighted estimator of  $\boldsymbol{\beta}$  can be written as

$$\hat{\boldsymbol{\beta}} = \sum_{i \in s} \sum_{k \in s_i} \mathbf{A}^{-1} \mathbf{x}_{ik} w_{ik} Y_{ik} = \sum_{i \in s} \mathbf{A}^{-1} \mathbf{X}_i^T \mathbf{W}_i \mathbf{Y}_i$$

with  $w_{ik}$  and  $Y_{ik}$  being the weight and dependent variable for unit  $(ik)$ ,  $s$  the sample of clusters,  $s_i$  the sample of units from cluster  $i$ , and

$\mathbf{X}_i$  = the  $m_i \times p$  matrix of explanatory variables,  $\mathbf{x}_{ik}$ 's, for the  $m_i$  sample units in cluster  $i$ ,  $i = 1, \dots, n$ ;

$\mathbf{W}_i$  = the  $m_i \times m_i$  diagonal matrix of survey weights;  
 $\mathbf{Y}_i$  = the  $m_i$ -vector of  $Y_{ik}$ 's, and  
 $\mathbf{A} = \sum_{i \in S} \mathbf{X}_i^T \mathbf{W}_i \mathbf{X}_i$ .

Pfeffermann et al. (1998) proposed the probability-weighted iterative generalized least squares (PWIGLS) estimator to obtain consistent estimates of the population variance parameters  $\sigma_U^2$  and  $\rho_U$ . The PWIGLS estimator, which assumes that the sampling probabilities for both stages  $\pi_i$  and  $\pi_{k|i}$ , or equivalently,  $w_i$  and  $w_{k|i}$ , are known, is adapted from the standard iterative generalized least squares (IGLS) by analogy with PML. Alternative inflation-type estimators using the two-level sample weights have also been considered (Longford 1995, Graubard and Korn 1996). However, Korn and Graubard (2003) later showed that these estimators can be badly biased when the sampling is informative. They proposed a new set of approximately unbiased estimators for variance components regardless of the sampling design. The limitation of these estimators is that they require the knowledge of second-order inclusion probabilities of the observations. In many surveys, analysts will not know the value of  $M_i$ ,  $w_i$ ,  $w_{k|i}$ , or the joint inclusion probabilities. If so, the only workable approach is to use a purely model based estimator

$$\hat{P} = \frac{1}{n} \sum_{i \in S} \frac{1}{m_i - 1} \sum_{k \in s_i} (e_{ik} - \bar{e}_i)^2$$

$$\hat{Q} = \sum_{i \in S} m_i (\bar{e}_i - \bar{e})^2 / (n-1)$$

$$\hat{D} = \left( m_+ - \sum_{i \in S} m_i^2 / m_+ \right) / (n-1),$$

where  $m_+ = \sum_{i \in S} m_i$ , and the residuals are calculated from the OLS estimator without using the sample weights. Using the estimators  $\hat{P}$ ,  $\hat{Q}$ , and  $\hat{D}$  we can formulate estimators as

$$\widehat{(1-\rho)\sigma^2} = \hat{P}$$

$$\widehat{\rho\sigma^2} = (\hat{Q} - \hat{P}) / \hat{D}$$

Another alternative is to use analysis of variance or restricted maximum likelihood methods, for instance, in SAS PROC VARCOMP or PROC MIXED.

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

$$v_M(\hat{\beta}) = \sum_s \mathbf{A}^{-1} \mathbf{X}_i^T \mathbf{W}_i \left( \widehat{(1-\rho)\sigma^2} \mathbf{I}_{m_i} + \widehat{\rho\sigma^2} \mathbf{1}_{m_i} \mathbf{1}_{m_i}^T \right) \mathbf{W}_i \mathbf{X}_i \mathbf{A}^{-1}.$$

This estimator is highly dependent on the working model and is not robust to departures from that model.

An alternative, sandwich estimator is consistent under a reasonably general variance specification. Consider the model:

$$E_M(Y_{ik}) = \mathbf{x}_{ik}^T \boldsymbol{\beta}$$

$$Cov_M(Y_{ik}, Y_{i'k'}) = 0 \quad i \neq i'. \quad (2)$$

Within a cluster, each pair of units could have a different correlation. The variance estimator will be derived using the cluster-level residuals and have the sandwich form. The vector of sample residuals for cluster  $i$  is  $\mathbf{e}_i = \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$ , and the residual for sample unit  $(ik)$  is  $e_{ik} = Y_{ik} - \mathbf{x}_{ik}^T \hat{\boldsymbol{\beta}}$ . If  $\mathbf{A}^{-1} = O(N^{-1})$ , and the sample sizes  $m_i$  are bounded, as the number of sampled PSUs becomes large, or  $n \rightarrow \infty$ ,  $E_M(\mathbf{e}_i \mathbf{e}_i^T) \cong V_M(\mathbf{Y}_i)$ , and consequently, the sandwich variance estimator is

$$v_W(\hat{\beta}) = \sum_{i \in S} \mathbf{A}^{-1} \mathbf{X}_i^T \mathbf{W}_i (\mathbf{e}_i \mathbf{e}_i^T) \mathbf{W}_i \mathbf{X}_i \mathbf{A}^{-1}.$$

Another useful variance estimator is the design-based linearization variance estimator. The linear approximation (Fuller 2002) of  $\hat{\boldsymbol{\beta}}$  is

$$\hat{\boldsymbol{\beta}} - \mathbf{B} \doteq \mathbf{A}_N^{-1} \sum_{i \in S} \sum_{k \in s_i} \mathbf{x}_{ik} w_{ik} (Y_{ik} - \mathbf{x}_{ik}^T \mathbf{B}) = \sum_{i \in S} \mathbf{z}_i$$

where  $\mathbf{z}_i = \mathbf{A}_N^{-1} \mathbf{X}_i^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{X}_i \mathbf{B})$ ,  $\mathbf{A}_N = \mathbf{X}_N^T \mathbf{X}_N$ , and  $\mathbf{B} = \mathbf{A}_N^{-1} \mathbf{X}_N^T \mathbf{Y}_N$ . Assuming the first-stage sample was selected with replacement, a design-based linearization estimator is given as

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

where  $\mathbf{z}_i^* = \mathbf{A}^{-1} \mathbf{X}_i^T \mathbf{W}_i \mathbf{e}_i = \mathbf{A}^{-1} \mathbf{X}_i^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})$  is a vector of  $p$  elements computed from PSU  $i$  and estimates  $\mathbf{z}_i$ . Note that  $\bar{\mathbf{z}}^* = n^{-1} \sum_s \mathbf{A}^{-1} \mathbf{X}_i^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}) = \mathbf{0}$ .

Then the model-based variance estimator  $v_W$  and the design-based variance estimator  $v_L$  would be approximately the same when the number of sampled clusters is large.

There are multiple ways to account for stratification in the modeling, depending on different model assumptions, but we would not cover this topic in this paper.

### 3. Identifying Single Influential Observations

As examined in the preliminary study (Li and Valliant 2006), the diagnostic tools were designed to measure the discrepancy in various statistics, such as regression coefficients and fitted values, between with and without potentially influential points in the model fitting. For complex survey data, diagnostics that incorporate variance estimators must account for complex sample designs. We use model-based variance estimators to derive the cutoff values for the diagnostic statistics in the light of distributional properties. Moreover, sandwich and linearization estimators can also be used to gain some protection against model failure and obtain design-based interpretations for the adapted statistics.

Leverages, which are the diagonal elements on the hat matrix, are helpful for detecting observations with outlying  $X$  values and sample weights, as addressed in Li and Valliant (2006). This statistic, as opposed to ones described below, does not involve variance estimates. Residuals, which are used to filter points with outlying  $Y$  values, usually are standardized with respect to  $\sqrt{MSE}$ . For clustered sampling and its corresponding model (1), we can divide  $e_{ik}$  by  $\hat{\sigma} = \sqrt{\frac{\sum_{i \in s} \hat{Q}_i - P}{P} D^{-1}}$ . If  $e_{ik}$  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  $\mu_0$ , then  $P\{|x - \mu_0| > \lambda\tau\} \leq 4/9\lambda^2$ , where  $\tau^2 \equiv \sigma^2 + (\mu - \mu_0)^2$ .

Under model (1) the residual has a symmetric distribution with its mode and mean at zero. The Gauss Inequality explains that the absolute value of a residual has 90% probability to be less than twice its standard deviation and 95% probability to be less than three times its standard deviation. If we rescale the residuals by a consistent estimate of  $\sigma$ , we can use either 2 as a loose cutoff or 3 as a strict one to identify outlying residuals, depending on an analyst's preference. Note that it is not feasible to define the distribution of residuals from a purely design-based point of view, even asymptotically because a residual refers to a single unit and is not a summation across units.

#### 3.1 DFBETAS

To measure the difference in each estimated coefficient after the  $(ik)$ th unit is deleted, we define

$$DFBETA_{ik} = \hat{\beta} - \hat{\beta}(ik) = \frac{\mathbf{A}^{-1} \mathbf{x}_{ik} e_{ik} w_{ik}}{1 - h_{ik,ik}},$$

where  $h_{ik,ik} = \mathbf{x}_{ik}^T \mathbf{A}^{-1} \mathbf{x}_{ik} w_{ik}$  is the  $k$ th diagonal element on the matrix  $\mathbf{H}_{ii} = \mathbf{X}_i \mathbf{A}^{-1} \mathbf{X}_i^T \mathbf{W}_i$  (see, e.g., Valliant, et al. 2000). Since

$$v_M(\hat{\beta}) = \hat{\sigma}^2 \sum_s \mathbf{C}_i \left[ (1 - \hat{\rho}) \mathbf{I}_{m_i} + \hat{\rho} \mathbf{1}_{m_i} \mathbf{1}_{m_i}^T \right] \mathbf{C}_i^T$$

where  $\mathbf{C}_i$  is a  $p \times m_i$  submatrix of  $\mathbf{C}$  defined as

$\mathbf{C}_i = \mathbf{A}^{-1} \mathbf{X}_i \mathbf{W}_i$  with  $(jk)$ th element  $c_{j,ik}$ , we also have

$$v_M(\hat{\beta}_j) = \hat{\sigma}^2 \sum_s \left( \sum_{k=1}^{m_i} c_{j,ik}^2 + \hat{\rho} \sum_{k \neq k'} c_{j,ik} c_{j,ik'} \right).$$

The constructed DFBETAS statistic is specified as

$$DFBETAS_{ik,j} = \frac{c_{j,ik} e_{ik} / (1 - h_{ik,ik})}{\sqrt{v_M(\hat{\beta}_j)}} = \frac{c_{j,ik}}{\sqrt{\sum_s \left( \sum_{k=1}^{m_i} c_{j,ik}^2 + \hat{\rho} \sum_{k \neq l} c_{j,ik} c_{j,il} \right)}} \cdot \frac{e_{ik}}{\hat{\sigma}} \cdot \frac{1}{1 - h_{ik,ik}}.$$

In order to define a cutoff for identifying extreme points, some simplifications are needed. If the  $\mathbf{X}$  variables and the sample weights  $\mathbf{W}$  are approximately equal for units across the clusters and the sample sizes within each cluster do not vary to a large degree, the first term in the above formula will be nearly the same as  $1/\sqrt{nm\bar{m}[1 + \hat{\rho}(\bar{m} - 1)]}$  with  $\bar{m} = n^{-1} \sum_s m_i$ . Note that

$1 + \hat{\rho}(\bar{m} - 1)$  is the estimated design effect. We propose that the cutoff value for DFBETAS statistics can be set as  $2/\sqrt{nm\bar{m}[1 + \hat{\rho}(\bar{m} - 1)]}$  or  $3/\sqrt{nm\bar{m}[1 + \hat{\rho}(\bar{m} - 1)]}$ . If there is only one unit within each sampled PSU, or  $\bar{m} = 1$ , the above cutoff boils down to  $2/\sqrt{n}$  or  $3/\sqrt{n}$  with  $n$  defined as the sample size, which corresponds to what we proposed in Li and Valliant (2006).

#### 3.2 DFFITS

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

$$DFFIT_{ik} = \hat{Y}_{ik} - \hat{Y}_{ik}(ik) = \frac{h_{ik,ik} e_{ik}}{1 - h_{ik,ik}}.$$

The variance of the predicted value is estimated as

$$v_M(\hat{Y}_{ik}) = \mathbf{x}_{ik}^T v_M(\hat{\boldsymbol{\beta}}) \mathbf{x}_{ik} = \hat{\sigma}^2 \sum_{i' \in s} \left( \sum_{k'=1}^{m_{i'}} h_{ik,i'k'}^2 + \hat{\rho} \sum_{k'' \neq k'}^{m_{i'}} h_{ik,i'k'} h_{ik,i'k''} \right)$$

Therefore, the DFFITS statistic is formulated as

$$DFFITS_{ik} = \frac{h_{ik,ik} e_{ik} / (1 - h_{ik,ik})}{\sqrt{v_M(\hat{Y}_{ik})}} = \frac{e_{ik}}{\hat{\sigma}} \frac{1}{\sqrt{\sum_{i' \in s} \left( \sum_{k'=1}^{m_{i'}} h_{ik,i'k'}^2 + \hat{\rho} \sum_{k'' \neq k'}^{m_{i'}} h_{ik,i'k'} h_{ik,i'k''} \right)}} \frac{h_{ik,ik}}{1 - h_{ik,ik}}$$

where  $h_{ik,i'k'} = \mathbf{x}_{ik}^T \mathbf{A}^{-1} \mathbf{x}_{i'k'} w_{i'k'}$  is an element of  $\mathbf{H}_{i'i'} = \mathbf{X}_i \mathbf{A}^{-1} \mathbf{X}_{i'}^T \mathbf{W}_{i'}$ . We can make approximations analogous to the ones used for DFBETAS in order to justify a cutoff for DFFITS. If  $\mathbf{X}$ ,  $\mathbf{W}$ , and  $m_i$  are similar across the clusters,

$$\sum_{i' \in s} \left( \sum_{k'=1}^{m_{i'}} h_{ik,i'k'}^2 + \hat{\rho} \sum_{k'' \neq k'}^{m_{i'}} h_{ik,i'k'} h_{ik,i'k''} \right) \approx [1 + \hat{\rho}(\bar{m} - 1)] h_{ik,ik}$$

The cutoff for the DFFITS statistic is determined to be  $2\sqrt{p/n\bar{m}[1 + \hat{\rho}(\bar{m} - 1)]}$  when  $\rho$  is appropriately estimated. We can also consider  $3\sqrt{p/n\bar{m}[1 + \hat{\rho}(\bar{m} - 1)]}$  as a less strict cutoff.

### 3.3 Modified Cook's Distance

If we assume the working model is (1), a quadratic statistic can be constructed as

$$ED_{ik} = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(ik))^T [v_M(\hat{\boldsymbol{\beta}})]^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(ik)) = \left( \frac{e_{ik}}{\hat{\sigma}} \right)^2 \frac{1}{(1 - h_{ik,ik})^2} w_{ik} \mathbf{x}_{ik}^T [\mathbf{X}^T \mathbf{W} \boldsymbol{\Phi} \mathbf{W} \mathbf{X}]^{-1} \mathbf{x}_{ik} w_{ik}$$

where  $\hat{\boldsymbol{\beta}}(ik)$  is the parameter estimate after deleting unit  $k$  in cluster  $i$ , the matrix  $\boldsymbol{\Phi}$  is block diagonal with 1 on the diagonal and  $\rho$  off the diagonal in each block (cluster), the dimension of block  $i$  is  $m_i \times m_i$ . If the number of units within each sampled PSU,  $m_i$ , is bounded,  $w_{ik} \mathbf{x}_{ik}^T [\mathbf{X}^T \mathbf{W} \boldsymbol{\Phi} \mathbf{W} \mathbf{X}]^{-1} \mathbf{x}_{ik} w_{ik} = O(n^{-1})$ , and the value of this expression is approximately equal to  $p[n\bar{m}(1 + \hat{\rho}(\bar{m} - 1))]^{-1}$  when the auxiliary variables  $\mathbf{X}$  and survey weights  $\mathbf{W}$  do not vary dramatically. Therefore, in the clustered sampling case we can compare

the square root of  $ED_{ik}$  with the cutoff value  $2\sqrt{p/n\bar{m}[1 + \hat{\rho}(\bar{m} - 1)]}$  or  $3\sqrt{p/n\bar{m}[1 + \hat{\rho}(\bar{m} - 1)]}$ .

Also, we can define a statistic named after the classical Cook's Distance, the Modified Cook's Distance, as

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

and compare it to 2 or 3.

Analysts can choose the diagnostic approaches and cutoff values in terms of different design features and model assumptions. The model-based variance estimators in the diagnostic statistics can always be replaced by the sandwich variance estimator and the linearization variance estimator to obtain protection against model misspecification.

### 4. Identifying Influential Groups of Observations

In large datasets the effect of groups of influential points can be masked when the entire dataset is used for model fitting. Atkinson and Riani (2000) introduced an effective and robust method of identifying such masked outliers, "the forward search", which seeks to divide the data into two parts, a large "clean" part and the outliers. Their emphasis, similar to DFBETA and Cook's distance, is on the change in parameter estimation once some of the data, including the outliers, have been removed. Here is a general description of how the forward search method may be modified and implemented in a single-stage survey sample.

- (1) Select a "clean" initial subset of size  $m$  from the sample, which is assumed not to include any outlier.
- (2) From the rest of  $n - m$  observations, add one observation at a time to construct a new subset of size  $m + 1$ , and calculate a key statistic which measures the change in regression parameters if this observation were removed from the subset. Cycle through all  $n - m$  observations to obtain  $n - m$  values of the key statistic.
- (3) Retain the observation with the minimum key statistic.
- (4) Repeat steps (2)-(3) until all observations are included in the regression.

As the algorithm proceeds sequentially through the points, outlying values will enter last and cause abrupt changes in the key statistic.

There are three important issues to consider for this algorithm to function appropriately. The first is the choice of the initial subset which is free of outliers and has a desirably small sample size. To avoid the inclusion of outliers in the starting subset, we may select points from the pool of observations which are not identified by any of the single-case deletion approaches, or keep a

group with the minimum median of squared residuals (LMS). Choosing the key statistic is the second important issue. The modified Cook’s Distance is a suitable choice because it summarizes the changes in all regression parameters and has stable performance. Other statistics can also be considered. The third issue is to determine the cutoff value for the key statistic. An analyst may simply use a fixed cutoff, such as 2 or 3. However, we suggest making a case-by-case judgment in which the analyst can account for the changing trends of the key statistic.

We also recommend that different initial subsets and various key statistics be applied to complete multiple searching processes so that we can confirm that the same group of outliers will enter into the subset at the last several steps. Moreover, the selection of the initial subset must consider the characteristics of the survey design, for example, clustering and stratification in order to produce correct estimates of regression parameters. Assuming a two-stage stratified clustering design, at each stage of model fitting, the set of units used needs to provide an estimate of the full population parameter. This implies that the initial set used for robust estimation must cover all strata and at least one PSU in each stratum.

### 5. Simulation

To evaluate the performance of the diagnostic approaches proposed and modified in Section 3 and 4, we conducted a simulation study and examined whether the methods of influence detection can be used to estimate the regression parameters better than the estimates that simply use all units. We generated a population in which the underlying model was known and then injected outlying observations. Thus, the correct “core” model is known, and it is possible to evaluate how well that model is estimated after identifying and deleting influential cases.

#### 5.1 Study Population and Sample Design

The population used in the simulation was created from the 1998 Survey of Mental Health Organizations (SMHO). The SMHO population has 875 observations and two auxiliary variables, number of beds and number of additions. First we kept 543 cases with number of beds between 10 and 300 and number of additions between 10 and 7000 as the “core” part of the study population. A  $\mathbf{Y}$  vector was then generated based on the two auxiliary variables using Gamma distributions  $Y_i \sim \text{Gamma}(s, a)$ ,

with shape parameter  $s = \sigma^2 / \mathbf{x}_i^T \boldsymbol{\beta}$  and scale parameter  $a = (\mathbf{x}_i^T \boldsymbol{\beta})^2 / \sigma^2$ ,  $\mathbf{x}_i$  is a vector including intercept, number of beds, and number of additions,

$\boldsymbol{\beta} = (5000, 80, 4)^T$ , and  $\sigma^2 = 8 \times 10^6$ . With these parameters,  $Y_i$  has a mean  $\mathbf{x}_i^T \boldsymbol{\beta}$  and a constant variance  $\sigma^2$ .

Five possible influential points were created and added to the “core” population. The  $\mathbf{X}$  values for the 5 outliers were generated from two uniform distributions. Number of beds was selected between 200 and 300 and number of additions was chosen between 4000 and 8000. The corresponding  $\mathbf{Y}$  were created using  $\mathbf{Y} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$ , where  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \tilde{\sigma}^2 \mathbf{I})$ ,  $\tilde{\boldsymbol{\beta}} = (500, 10, 1)^T$ , and  $\tilde{\sigma}^2 = 10^3$ .

Therefore, the study population consists of three variables and has a size of 548. Figure 1 displays the positions of the outlying units with respect to the “core” population, and illustrates that the generated outliers are likely to pull the potential “core” regression line downwards. Since our goal of inference is to develop procedures that permit good estimates of parameters for a model that fits reasonably well for most of a finite population, we used the OLS estimates on the “core” population to be the “core” parameters. Table 1 shows the parameter estimates from the regression of  $\mathbf{Y}$  on number of beds and number of additions based on the “core” population and the full population, respectively. The estimated coefficients based on the 543 “core” cases are very close to the “core” model parameters. However, when the generated outliers were included in the regression, the slope estimates substantially decreased to 56.72 and 3.5.

Table 1. Parameter Estimations Based on “Core” Population and Full Population with 5 Outliers.

Independent Variables	Finite Population Parameters		
	Underlying Core Model	Core	Full
Intercept	5000	5057 (239)	7099 (363)
Beds	80	76.01 (2.48)	56.72 (3.78)
Additions	4	4.09 (0.09)	3.50 (0.15)

For simplicity, we selected samples using a one-stage probability proportional to size (PPS) design and the measure of size being the 0.85 power of number of beds. The created outliers in the population are associated with relatively large number of beds so that they are more likely to be selected and, if selected, have smaller sample weights. In each sample, 100 units were drawn without replacement. Sample weights were calculated based on the selection probabilities.

#### 5.2 Diagnostic Scheme and Regression

Besides the comparison between the estimates from full samples and reduced samples, we are also interested in the difference between the OLS and the SW diagnostics.

Therefore, both the OLS and the SW diagnostics will be employed for each selected sample, and the diagnostic methods include DFBETAS, DFFITS, and Cook's Distance. For the SW diagnostic statistics, we used linearization variance estimators where needed, and a more strict criterion, 2, was used to construct cutoffs. When we utilized DFBETAS statistics to detect influential units, we examined units with extreme DFBETAS for both auxiliary variables. In all, based on each selected sample, we were able to create 6 reduced subsamples. The OLS and the SW regressions were run on full samples and the corresponding regressions were run on reduced subsamples.

### 5.3 Summary Statistics

The entire sampling, diagnostic, and regression process was repeated  $i=1, \dots, 5000$  times in the simulation. Summary statistics across the simulation include:

(1) Average number of identified points and average number of identified outliers created in the constructed population.

(2) The average parameter estimates and their relative biases compared to the finite population "core" model. The relative bias was estimated by  $relbias(\hat{\beta}) = \frac{(\bar{\hat{\beta}} - \beta)}{\beta}$ , where  $\bar{\hat{\beta}} = \sum_i \hat{\beta}^{(i)} / 5000$ ,  $\hat{\beta}^{(i)}$  is the estimate of the parameter vector from sample  $i$ , and  $\beta = (5056.62, 76.01, 4.09)^T$  is the finite population "core" parameter vector.

(3) The estimated SEs of model parameter estimates  $se(\hat{\beta}) = \sqrt{\sum_i v(\hat{\beta}^{(i)})} / 5000$  as compared to the empirical

SEs  $Se(\hat{\beta}) = \sqrt{\sum_i (\hat{\beta}^{(i)} - \bar{\hat{\beta}})^2} / 5000$ , where  $v(\hat{\beta}^{(i)})$  is

the estimated variance of  $\hat{\beta}^{(i)}$  in the  $i$ th sample.

(4) The percentages of the 95% confidence intervals  $\hat{\beta}^{(i)} \pm 1.96 \sqrt{v(\hat{\beta}^{(i)})}$  that include the finite population "core" parameters.

### 5.4 Simulation Results

Table 2 reports the average number of units and average population outliers that were identified by each diagnostic method, either OLS or SW. Out of the 2.9 population outliers that were sampled on average, all of them can be recognized using the OLS and the SW diagnostic techniques. The results of the SW diagnostics showed that some points, which were not labelled as outlying in

the population, but were associated with moderate or large sample weights, could still play a crucial role in the regression estimation and be identified as influential. Those points were not counted as correctly identified outliers. But, we expect that the elimination of them would perceptibly change the regression estimates.

Table 2. Number of Influential Observations Identified and Population Outliers Identified.

Diagnostic Approaches	Average # of Outliers Identified	Average # of Pop Outliers Identified
OLS DFBETAS	9.3	2.9
SW DFBETAS	5.9	2.9
OLS DFFITS	6.2	2.9
SW DFFITS	10.7	2.9
OLS Cook's D	6.0	2.9
SW Cook's D	6.7	2.9
Average # of Outliers Sampled: 2.9		

The relative biases across the iterations, listed in Table 3, are good indicators to gauge the effectiveness of the diagnostic methods. (Lines for DFBETAS and DFFITS in Tables 3-6 refer to cases where the statistics for either beds or adds identified outliers.) Diagnostic approaches are useful to reduce the biases in both the OLS and the SW full sample estimates with respect to the core parameters. The relative biases for the slope estimates after applying OLS diagnostics ranged from -6.1 to -1.3% and from -3.6 to -0.9% using the SW diagnostics.

Table 3. Relative Biases.

	RelBias(%)		
	Intercept	Beds	Adds
Full Sample OLS	110.1	-44.7	-36.6
Full Sample SW	41.0	-24.6	-15.5
OLS DFBETAS	9.5	-6.1	-2.1
SW DFBETAS	5.4	-0.9	-2.6
OLS DFFITS	5.9	-3.6	-1.4
SW DFFITS	8.6	-3.6	-2.4
OLS Cook's D	5.7	-3.4	-1.3
SW Cook's D	7.8	-1.9	-2.9

Besides biases, it is also interesting to examine the real coverage rates of the confidence intervals constructed from the parameter estimates and their estimated standard errors at some nominal confidence level, which are reported in Table 4. The confidence intervals based on the OLS full sample estimates have extremely low chances to cover the core model parameters. When survey weights were accounted for, the coverage rates increased to more than 70%, but still 25% short of the nominal level. After the influential observations were successfully recognized and excluded from the regressions, the real coverage rates rose to about 90% for the slope parameters. Note that the OLS intervals had

uniformly higher coverage rates than SW intervals after applying the diagnostics.

Table 3 and 4 show that sometimes the SW estimates were less biased but had smaller coverage rates than the OLS estimates. Therefore, it is helpful to understand this problem by investigating the standard errors of the estimated coefficients. From Table 5 we conclude that some of the standard errors were underestimated for the regressions on the reduced samples. The likely reason of underestimating the SEs for SW regressions is that the variation in the number of observations used in the regressions was not accounted for. This phenomenon of underestimation is similar to what occurs with standard error estimates in stepwise regression (Hurvich and Tsai, 1990; Zhang 1992). For OLS regressions, including unidentified outliers in the model fitting can cause smaller estimated SEs than what they should be. For SW regressions, underestimation can be more severe if too many observations with large sample weights are detected as influential and eliminated from the sample.

Table 4. Coverage Rates of 95% CIs.

	Real Coverage Rate of the 95% CI		
	Intercept(%)	Beds(%)	Adds(%)
Full Sample OLS	4	11	13
Full Sample SW	73	71	78
OLS DFBETAS	90	89	95
SW DFBETAS	87	91	89
OLS DFFITS	93	93	97
SW DFFITS	80	86	88
OLS Cook's D	94	93	97
SW Cook's D	85	91	89

Table 5. Ratios of Estimated SEs to Empirical SEs.

	Real Coverage Rate of the 95% CI		
	Intercept(%)	Beds(%)	Adds(%)
Full Sample OLS	0.74	0.92	0.71
Full Sample SW	1.24	1.27	1.12
OLS DFBETAS	1.01	1.11	1.08
SW DFBETAS	0.84	0.89	0.93
OLS DFFITS	1.00	1.05	1.13
SW DFFITS	0.75	0.84	0.88
OLS Cook's D	1.00	1.05	1.13
SW Cook's D	0.82	0.89	0.92

### 5.5 Simulation Results: Forward Search

In order to investigate the masked effect among outlier we designed another simulation in which 25 outliers were created using the same approach. Even if applying diagnostics can still improve the estimates, the results are far from being acceptable, as listed in Table 6. The unsatisfactory performance of the diagnostics is caused by the masked effect among outliers and can be resolved using the forward search method.

A new simulation involving an adapted forward search was set up with 1000 runs. Twenty units that are least likely to be detected by any single-case deletion method were chosen as the initial subsets. According to the results from a few pilot studies, we determined to use a cutoff value of 2.3 and define the observations as outliers if they have modified Cook's Distance larger than 2.3.

The new results show in Table 6 that the bias of the intercept dropped to 8.4%, and the biases of the estimated slopes decreased to -4.5% and -4.2%, respectively. The real coverage rates of the 95% CIs rise to 75%-80%. The standard errors were still underestimated. The negatively biased SE estimates are the main reason for undercoverage of the confidence intervals when the forward search is used, rather than bias in the parameter estimates. By avoiding the masked effect among the outliers, the forward search method identifies the influential group more correctly. We expect the parameter estimates would be even less biased if we exercise more control over the selection of the initial subset and where to drop the line between the "clean" part and the outliers.

## 6. Conclusion

With the incorporation of survey weights and design features, we constructed survey weighted diagnostic statistics in a way similar to the conventional OLS diagnostics. Survey weighted diagnostics may identify different points than OLS diagnostics as being influential. An observation with moderate  $Y$  and  $x$  values may not be identified as influential by OLS approaches, but may be recognized as influential by SW methods if it is assigned an extreme sample weight. Techniques based on single-case deletion may not function effectively when some outliers mask the effects of others. The modified forward search method is a partial solution to this problem since it can successfully identify the influential group and avoid masked effect among outliers. The diagnostics can serve as a guide to which points may be unusual. However, a diligent analyst should examine these points in detail to decide whether they are data entry errors, legitimate values that do not follow a core model, or can be explained in some other way like having extreme weights.

Once influential observations or group are caught, a natural but not unique remedy is to remove them from the regression. Dropping influential points and refitting models may produce different parameter estimates from full sample estimates and therefore affect inferences about the population. If too many or too few outliers are identified than appropriate, it can cause incomplete correction of bias, underestimation of variance, and as a

result, the coverage rate of constructed confidence intervals will be less than nominal.

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

**Acknowledgements:** This paper is based upon work supported by the National Science Foundation under Grant No. 0617081.

### References

- Atkinson, A. C., and Riani, M. (2000), *Robust Diagnostic Regression Analysis*, New York: Springer-Verlag.
- Binder, D. A. (1983), "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review*, **51**, 279-292.
- 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.
- Elliott, M. (2007), "Bayesian weight trimming for generalized linear regression models," *Survey Methodology*, **33**, 23-34.
- Fuller, W. A. (2002), "Regression estimation for survey samples", *Survey Methodology*, **28**, No. 1, 5-23.
- Graubard, B. I., and Korn, E. L. (1996), "Modelling the sampling design in the analysis of health surveys," *Statistical Methods in Medical Research*, **5**, 263-281.
- 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: Wiley.
- Korn, E. L., and Graubard, B. I. (2003), "Estimating variance components by using survey data," *Journal of Royal Statistical Society*, **65**, Part 1, 175-190.
- Kott, P. S. (1991), "A model-based look at linear regression with survey data", *American Statistician* **45**: 107-112.
- Li, J. (2007). *Regression Diagnostics for Complex Survey Data: Identification of Influential Observations*. Ph.D. 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.
- 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, A*, **148**: 268-278.
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998), "Weighting for unequal selection probabilities in multilevel models," *Journal of the Royal Statistical Society, Series B, Methodological*, **60**, 23-40.
- Pukelsheim, F. (1994), "The three sigma rule," *The American Statistician*, **48**, 88-91.
- Skinner, C. J., Holt, D. and Smith, T. M. F. (eds.) (1989), *Analysis of Complex Surveys*, New York: Wiley.
- Valliant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, New York: Wiley.
- Zhang, P. (1992), "Influence after variable selection in linear regression models," *Biometrika*, **79**, 741-746.

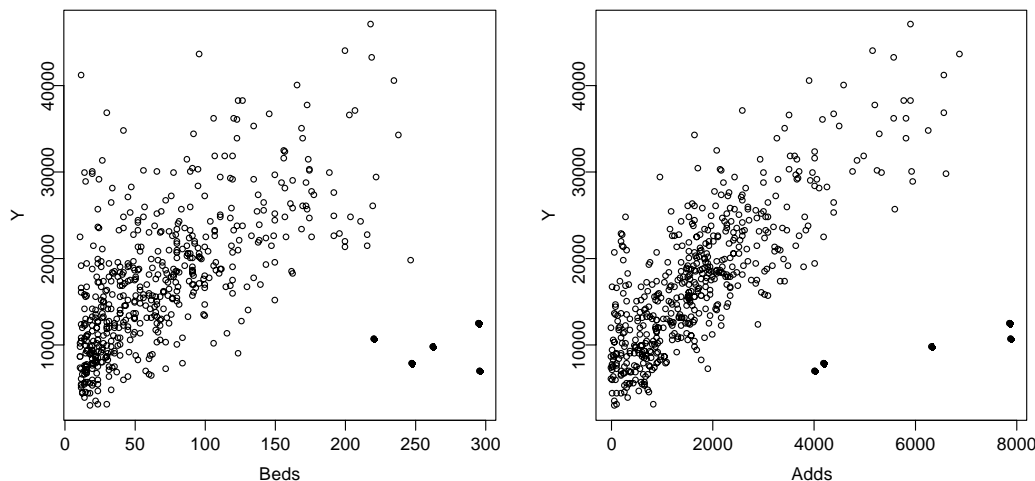


Figure 1. Plots of Y versus Auxiliary Variables Including 5 Generated Outliers shown as black dots in the lower right of each plot.