ABSTRACT

Title of Dissertation: REGRESSION DIAGNOSTICS FOR COMPLEX SURVEY DATA: IDENTIFICATION OF INFLUENTIAL OBSERVATIONS.

Jianzhu Li, Doctor of Philosophy, 2007

Dissertation Directed By: Professor Richard Valliant
Joint Program in Survey Methodology

Discussion of diagnostics for linear regression models have become indispensable chapters or sections in most of the statistical textbooks. However, survey literature has not given much attention to this problem. Examples from real surveys show that sometimes the inclusion and exclusion of a small number of the sampled units can greatly change the regression parameter estimates, which indicates that techniques of identifying the influential units are necessary. The goal of this research is to extend and adapt the conventional ordinary least squares influence diagnostics to complex survey data, and determine how they should be justified.

We assume that an analyst is looking for a linear regression model that fits reasonably well for the bulk of the finite population and chooses to use the survey weighted regression estimator. Diagnostic statistics such as DFBETAS, DFFITS, and modified Cook's Distance are constructed to evaluate the effect on the regression coefficients of deleting a single observation. As components of the diagnostic statistics,

the estimated variances of the coefficients are obtained from design-consistent estimators which account for complex design features, e.g. clustering and stratification. For survey data, sample weights, which are computed with the primary goal of estimating finite population statistics, are sources of influence besides the response variable and the predictor variables, and therefore need to be incorporated into influence measurement. The forward search method is also adapted to identify influential observations as a group when there is possible masked effect among the outlying observations.

Two case studies and simulations are done in this dissertation to test the performance of the adapted diagnostic statistics. We reach the conclusion that removing the identified influential observations from the model fitting can obtain less biased estimated coefficients. The standard errors of the coefficients may be underestimated since the variation in the number of observations used in the regressions was not accounted for.

REGRESSION DIAGNOSTICS FOR COMPLEX SURVEY DATA: IDENTIFICATION OF INFLUENTIAL OBSERVATIONS

By

Jianzhu Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Dr. Richard Valliant, Chair
Dr. Barry Graubard
Dr. Partha Lahiri
Dr. Stephen Miller
Dr. Paul Smith

Dedication

This Dissertation is dedicated to my parents, Ruoxin Li and Guiqin Liang, and to my daughter, Yifan Li.

Acknowledgements

I would like to express my sincerest gratitude to my advisor, Professor Richard Valliant, for his guidance and support throughout my education and my research at University of Maryland. This dissertation would not be possible without his help.

I would like to thank Dr. Barry Graubard, Dr. Partha Lahiri, Dr. Stephen Miller, and Dr. Paul Smith for agreeing to be members of dissertation committee and for helping me to clarify my research problem.

I wish to thank Professor Roger Tourangeau for providing me support and encouragement to strive for this degree.

I would also like to thank Rupa Jethwa and Adam Kelley for their administrative and technical support, Jill Dever and other fellow members of the JPSM PhD cohort for their encouragement and help.

This Dissertation is based upon work supported by the National Science Foundation under Grant No. 0617081.

Table of Contents

Dedication ii
Acknowledgements iii
Table of Contents iv
List of Tables vi
List of Figures viii
Chapter 1: Introduction 1
1.1 Literature Review 1
1.2 Uses of Survey Data 2
1.3 The Subject of This Dissertation 3
Chapter 2: Linear Regression Analysis 5
2.1 Traditional Linear Regression Model 5
2.2 Linear Regression for Complex Survey Data 5
Chapter 3: Identification of Single Influential Observations 9
3.1 Introduction 9
3.2 Basic Idea in Influence Assessment 10
3.3 Sources of Influence in Survey Data 11
3.4 Review of Traditional Techniques 12
3.4.1 Leverages and Residuals 12
3.4.2 Influence on Regression Coefficients: DFBETA and DFBETAS 13
3.4.3 Influence on Fitted Values: DFFIT and DFFITS 14
3.4.4 Cook's Distance 14
3.5 Variance Estimation Methods for Complex Survey Data 16
3.5.1 Asymptotic Framework 17
3.5.2 Variance Estimation for Single-Stage Sampling With Replacement 19
3.5.3 Variance Estimation for Multistage Sampling Design 23
3.6 Adaptations of Traditional Techniques to Regression on Complex Survey Data 31
3.6.1 Residuals and Leverages 31
3.6.2 DFBETAS 38
3.6.3 DFFITS 41
3.6.4 Distance Measure (Extended and Modified Cook's Distance) 43
3.6.5 Discussion 48
Chapter 4: Identification of Influential Groups of Observations 50
4.1 Multiple-Case Deletion 50
4.2 Deletion of Specific Characteristic Groups 55
4.3 Forward Search 56
4.3.1 Introduction 56
Chapter 5: Application of Diagnostic Techniques for Influence Analysis 61
5.1 Introduction 61
5.2 Identifying Single Influential Observations: Case Study 1 62
5.2.1 Summary of SMHO Data Set 62

5.2.2 Parameter Estimation 64
5.2.3 Diagnostics by Leverages and Residuals 65
5.2.4 Diagnostics by DFBETAS 70
5.2.5 Diagnostics by DFFITS and Modified Cook's Distance 76
5.2.6 Discussion 79
5.3 Identifying Single Influential Observations: Case Study 2 80
5.3.1 Summary of NHANES Data Set 80
5.3.2 Diagnostic Results 82
5.4 Simulation 86
5.4.1 Description of Study Population and Sample Design 86
5.4.2 Diagnostic Scheme and Regression 90
5.4.3 Summary Statistics 90
5.4.4 Simulation Results 91
5.4.5 Possible Masked Effect among Outliers 98
5.4.5 Discussion 104
5.5 Case Studies Revisited: Forward Search Method 105
5.5.1 Case Study 1 Revisited: SMHO data 106
5.5.2 Case Study 2 Revisited: NHANES data 112
5.6 Simulation Revisited: Forward Search Method 116
Chapter 6 Conclusion 118
References 121

List of Tables

Table 5.1. Quantiles of Variables in SMHO Regression. 63
Table 5.2. OLS and SW Parameter Estimates of SMHO Regression of Expenditures on Beds and Additions. 64
Table 5.3. OLS and SW Parameter Estimates after Deleting Observations with Large Leverages from SMHO Regression. 68
Table 5.4. OLS and SW Parameter Estimates after Deleting Observations with Large Residuals from SMHO Regression. 68
Table 5.5. OLS and SW Parameter Estimates after Deleting Observations with Large DFBETAS of Beds for SMHO Data. 74
Table 5.6. OLS and SW Parameter Estimates after Deleting Observations with Large DFBETAS of Adds for SMHO Data. 74
Table 5.7. OLS and SW Parameter Estimates after Deleting Observations with Large DFBETAS of either Beds or Adds for SMHO Data. 74
Table 5.8. OLS and SW Parameter Estimates after Deleting Observations with Large DFFITS for SMHO Data. 78
Table 5.9. OLS and SW Parameter Estimates after Deleting Observations with Large Modified Cook's Distance for SMHO Data. 78
Table 5.10. Quantiles of Variables in NHANES Regression of Systolic Blood Pressure on Age, BMI, and Blood Lead. 81
Table 5.11. OLS and SW Parameter Estimates from NHANES Regression. 81
Table 5.12. Number of Outliers Identified and Associated Weight Ranges for NHANES Data. 85
Table 5.13. Estimated Slopes of BMI from Full Sample and Reduced Samples by Different Diagnostic Approaches for NHANES Data. 86
Table 5.14. Parameter Estimations Based on "Core" Population and Full Population with 5 Outliers. 89
Table 5.15. Number of Influential Observations Identified and Correctly Identified in Population with 5 Outliers. 92
Table 5.16. Average Parameter Estimates and Relative Biases in Population with 5 Outliers. 93
Table 5.17. Coverage Rates of 95% Confidence Intervals in Population with 5 Outliers. 95
Table 5.18. Empirical and Estimated Standard Errors of Parameter Estimates in Population with 5 Outliers. 97
Table 5.19. Parameter Estimation Based on Population with 25 Generated Outliers. 99
Table 5.20. Number of Influential Observations Identified and Correctly Identified in Population with 25 Outliers. 100
Table 5.21. Average Parameter Estimates and Relative Biases in Population with 25 Outliers. 101
Table 5.22. Coverage Rates of 95% Confidence Intervals in Population with 25 Outliers. 102

Table 5.23. Empirical and Estimated Standard Errors of Parameter Estimates in Population with 25 Outliers. 103
Table 5.24. Parameter Estimates of SMHO Regression after Influential Group Identifed by Forward Search was Deleted. 111
Table 5.25. Parameter Estimates of NHANES Regression after Influential Group Identified by Forward Search was Deleted. 115
Table 5.26. Summary Statistics for Simulation using Forward Search Method. 117

List of Figures

Figure 5.1. OLS and SW residuals versus Two Auxilary Variables for SMHO Data. 65
Figure 5.2. Leverage and Residual Diagnostic Plots for SMHO Data 67
Figure 5.3. Fitted Values Plots After Applying Leverage and Residual Diagnostics to SMHO Data. 69
Figure 5.4. DFBETAS Plots for SMHO Data. 71
Figure 5.5. Scatterplots with OLS (top) and SW (bottom) Smoothing for SMHO Data 72
Figure 5.6. OLS and SW Added Variable Plots for SMHO Data 73
Figure 5.7. Fitted Values Plots After Applying DFBETAS Diagnostics to SMHO Data. 75
Figure 5.8. DFFITS and Modified Cook's Distance Plots for SMHO Data. 77
Figure 5.9. Fitted Values Plots After Applying DFFITS and Cook's D Diagnostics to SMHO Data. 79
Figure 5.10. Bubble Plots of Systolic Blood Pressure versus Three Auxiliary Variables for NHANES Data. 82
Figure 5.11. OLS and SW residuals versus Three Auxilary Variables for NHANES Data. 82
Figure 5.12. Leverage and Residual Plots for NHANES Data 84
Figure 5.13. DFBETAS Plot and Added Variable Plots of BMI for NHANES Data. ... 84
Figure 5.14. DFFITS Plot and Modified Cook's Distance Plot 85
Figure 5.15. Scatter Plot for Regression with One Predictor Variable Illustrating Outlying Cases. 88
Figure 5.16. Plots of Y versus Auxiliary Variables Including 5 Generated Outliers. 89
Figure 5.17. Dot Plot of Average Parameter Estimates and Relative Biases for OLS (+) Regressions and SW (.) Regressions in Population with 5 Outliers 94
Figure 5.18. Plots of Y versus Auxiliary Variables Including 25 Generated Outliers... 98
Figure 5.19. Plots of Single-Case Deletion Based Modified Cook's Distance from Forward Search with Two Different Initial Subsets in SMHO Data 108
Figure 5.20. Plots of Multiple-Case Deletion Extended Cook's Distance from Forward Search with Two Different Initial Subsets in SMHO Data. 108
Figure 5.21. Plots of MDFFIT from Forward Search with Two Different Initial Subsets in SMHO Data 109
Figure 5.22. Plots of Parameter Estimates from Forward Search with Two Different Initial Subsets in SMHO Data 109
Figure 5.23. Scatterplots with SW Scatterplot Smoothing and SW Added Variable Plots with Dark Bubbles Symbolizing Influential Points Identified by Forward Search for SMHO Data. 111
Figure 5.24. Plots of Single-Case Deletion Modified Cook's Distance from Forward Search with Two Different Initial Subsets for NHANES data. 113
Figure 5.25. Plots of Multiple-Case Deletion extended Cook's Distance from Forward Search with Two Different Initial Subsets for NHANES data. 113
Figure 5.26. Plots of MDFFIT from Forward Search with Two Different Initial Subsets

for NHANES data. 114
Figure 5.27. Plots of Estimated Slope of BMI from Forward Search with Two Different Initial Subsets for NHANES data. 115
Figure 5.28. Scatterplot of Systolic Blood Pressure versus BMI with Scatterplot Smoothing and Added Variable Plot of BMI for NHANES data..... 115

Chapter 1: Introduction

1.1 Literature Review

Several decades have passed since linear regression analysis became a widely employed statistical methodology that utilizes the relation between quantitative response and quantitative and qualitative covariates to make predictions and inferences. Regression attempts to model the relationship between two or more variables by fitting a linear equation to observed data. When a regression model is considered for an application, researchers and analysts usually are not certain in advance whether a particular form of model is appropriate, especially with social science or epidemiological data. It is therefore natural to raise questions before making inferences based on the particular data at hand. A general question is: what type of model is appropriate – linear or nonlinear? A more specific question is whether the fitted model is unduly affected by unusual points. If so, what features of the data explain this affect? Do collinear relationships exist among the data series used as predictors? Do such problems degrade the parameter estimation? Diagnostic techniques were gradually developed to find problems in model-fitting and to assess the quality and reliability of regression estimates. These concerns turned into an important area in regression theory intended to explore the characteristics of a fitted regression model for a given data set.

Discussion of diagnostics for linear regression models are often indispensable chapters or sections in most of the statistical textbooks on linear models. One of the most influential books on the topic was *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* by Belsley, Kuh, and Welsch (1980). Diagnostic statistics are also included as standard options in many statistical packages, for instance, SAS®, SPSS®, Stata®, and R®, and are now readily available to analysts who want to diagnose influential points, detect collinearity, and more.

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. The diagnostic tools provided by current,

popular software are generally based on ordinary or weighted least squares (OLS or WLS) regression and do not account for stratification, clustering, and survey weights that are features of data sets collected in complex sample surveys. The OLS/WLS diagnostics can mislead users either because survey weights are ignored, or the variances of model parameter estimates are estimated incorrectly by the standard procedures. Hence, the goal of this research is to adapt and extend some of the standard regression diagnostics to the survey setting, and, where necessary, develop new ones.

Survey literature has not given 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, et. al. (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. Chambers (1986), Gwet and Rivest (1992), Welsh and Ronchetti (1998), and Duchesne (1999) conduct research on outlier robust estimation techniques for totals. 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.

1.2 Uses of Survey Data

The application of conventional techniques to survey data becomes less straightforward because of features of complex sampling designs like stratification, clustering, and weights. Will standard diagnostic techniques still be useful after some modifications? How should we deal with the survey weights associated with each sampled unit? The use of survey data will be reviewed before we try to answer these questions.

The uses of surveys can be roughly divided into two categories: analytic and descriptive (Skinner, Holt and Smith, 1989). Descriptive uses of surveys usually involve the estimation of summary measures like means, totals, or quantiles of a finite

population based on samples taken according to a specific design. Traditionally, the use of models is incidental in design-based sampling because inferences are made about the population with respect to the randomization distribution of the samples. Curtailing the effects of unusual cases on the estimation of totals, means, and other descriptive statistics is done in the randomization approach by weight trimming or modification (e.g. see Potter 1990, Hidiroglou and Srinath 1981) or other informal methods. The prediction approach to survey sampling (Valliant, Dorfman, and Royall 2000) is an alternative way of making inferences about finite population parameters. This approach borrows strength from models established on the observed units and tries to accurately predict the unobservables in the population, and therefore is referred to as model-dependent. The quantities predicted are random variables whose realizations depend on fixed but unknown model parameters. Thus, the properties of estimators heavily rely on the quality of the model. Chambers (1996) proposes a modified method of linear regression-based case-weighting intended to ensure model misspecification robustness.

In contrast to the design-based approaches, analytic uses of surveys have essential involvement of model-building because investigators are interested in the properties (often causal relationships) of a wider "superpopulation" that the sampled population represents (Graubard and Korn, 2002). Population units are regarded not as fixed values but as the realizations of random variables whose distributions can be modeled using available information. It should be noted that clustering and stratification in the population, which are also reflected in a complex sampling design, cause the violation of standard model assumptions requiring independent and identical distribution of model errors. This makes the interpretations of the stochastic components of the model more complicated. Since models play a crucial role in the analysis of survey data, the diagnostics of model adequacy need to be carefully justified.

1.3 The Subject of This Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 will introduce the linear regression estimators with and without survey weights. The former is derived from the pseudo maximum likelihood method and used for the analysis of survey data,

while the latter is based on the traditional linear models and can be obtained by applying the ordinary least squares approach. The comparison of the two will shed light on the possible differences between conventional regression diagnostics and those for survey data. Chapter 3 and Chapter 4 will discuss influence assessment which is the core issue to be studied in this thesis. Chapter 3 will focus on the identification of individual influential observations. After reviewing some traditional techniques based on single-case deletion methods, diagnostic statistics such as DFBETAS, DFFITS, and Cook's Distance will be modified and adapted to the survey setting. Chapter 4 will use the same research methodology but will be devoted to locating influential groups. The forward search approach will be described and extended to survey design involving stratification and clustering. Chapter 5 consists of the application of newly-adapted statistics and approaches to real survey data and simulated data. Analysis will be given on the effectiveness of identifying influential individual observations or groups of observations. This study will conclude in Chapter 6 with a summary of limitations of the research and suggestions for future research to advance this work.

The new contributions in this dissertation are the adapted and modified diagnostic approaches which will be described in Chapter 3 and Chapter 4. The development of these methods allows us to conduct influence analysis on linear regression using complex survey data.

Chapter 2: Linear Regression Analysis

2.1 Traditional Linear Regression Model

Generally, for a linear regression under the nonsurvey setting, the model is formulated as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

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 $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$ are statistically independent error terms which are distributed with zero mean and constant variance σ^2. Hence, the Gauss-Markov theorem states that the least squares estimators are unbiased and have minimum variance among all unbiased linear estimators. The Ordinary Least Squares (OLS) estimator of parameter vector $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. If, in addition, the model errors are normally distributed, $\mathbf{b}$ is also the maximum likelihood estimator.

2.2 Linear Regression for Complex Survey Data

Parameter estimators in linear regression using complex survey data are derived from the Pseudo Maximum Likelihood (PML) approach, outlined by Skinner et al. (1989), following ideas of Binder (1983). The basic idea of this approach is that we could compute the likelihood and achieve consistent estimation by maximizing the likelihood if all population units were observed. Suppose that the underlying structural model is a fixed-effects linear model:

$$Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \text{ind } N(0, v_i\sigma^2) \tag{2.2}$$

where ε_i's are independently normally distributed with mean 0 and variance $v_i\sigma^2$,

which is known except for the constant σ^2. The likelihood for $\boldsymbol{\beta}$ is

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

where s is the set of sample units and $f\left(Y_i; \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2\right)$ is the normal density with mean $\mathbf{x}_i^T \boldsymbol{\beta}$ and variance $v_i \sigma^2$. If the full population were in the sample, the log-likelihood would be $\log L(\boldsymbol{\beta}) = \sum_{i \in U} f\left(Y_i; \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2\right)$. From this, the full population estimation equations are $\sum_{i \in U} \frac{\partial \log\left[f\left(Y_i; \boldsymbol{\beta}\right)\right]}{\partial \boldsymbol{\beta}} = \mathbf{0}$. These estimation equations are a type of finite population total for which a survey weighted estimator can be constructed. Thus, the Pseudo Maximum Likelihood Estimator (PMLE) of $\boldsymbol{\beta}$ is the solution to the set of estimation equations $\sum_{i \in s} w_i \frac{\partial \log\left[f\left(Y_i; \boldsymbol{\beta}\right)\right]}{\partial \boldsymbol{\beta}} = \mathbf{0}$, 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 PMLE to account for an informative design in which sample distribution of the 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

$$\sum_{i \in s} w_i \frac{Y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{v_i} \mathbf{x}_i = \mathbf{0} \quad \text{or} \quad \mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) = \mathbf{0}$$

with $\mathbf{V} = diag\left(v_1, ..., v_n\right)$ and $\mathbf{W} = diag\left(w_1, ..., w_n\right)$. These equations can be solved explicitly as $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{Y}$.

The regression estimator $\hat{\boldsymbol{\beta}}$ which incorporates the sample weights $\mathbf{W}$ is approximately design unbiased for the finite population parameter $\mathbf{B} = \left(\mathbf{X}_N^T \mathbf{V}_N^{-1} \mathbf{X}_N\right)^{-1} \mathbf{X}_N^T \mathbf{V}_N^{-1} \mathbf{Y}_N$, where $\mathbf{Y}_N = \left(Y_1, ..., Y_N\right)^T$, $\mathbf{V}_N = diag\left(v_1, ..., v_N\right)$, and $\mathbf{X}_N^T = \left(\mathbf{x}_1, ..., \mathbf{x}_N\right)$ (Särndal, Swensson, and Wretman, 1992). Approximate design

unbiasedness of $\hat{\boldsymbol{\beta}}$ means that its expectation over repeated sampling is approximately $\mathbf{B}$ assuming that the weights w_i's are constructed to produce design-unbiased estimates of finite population totals. From the model-based perspective, this estimator is also unbiased for the superpopulation slope $\boldsymbol{\beta}$ in model (2.2), regardless of whether $\mathbf{V}$ is specified correctly or not. When the population is large, the finite population parameter $\mathbf{B}$ should be close to the model parameter $\boldsymbol{\beta}$ if model is correctly specified, and therefore a design-based estimator of $\mathbf{B}$ should also estimate $\boldsymbol{\beta}$. If we assume $\mathbf{V}=\mathbf{I}$, model (2.2) reduces to (2.1) and the parameter estimator $\hat{\boldsymbol{\beta}}$ will consequently take the form of $\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$. This estimator will be referred as Survey Weighted (SW) estimator in the following discussion and is the one usually computed by software packages that handle survey data. Note that results for $\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{Y}$ can be obtained by replacing $\mathbf{W}$ in the SW estimator by $\mathbf{W}^*=\mathbf{W}\mathbf{V}^{-1}$.

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 or weighted least squares estimator that uses only $\mathbf{V}^{-1}$ (not $\mathbf{W}$) can be used to infer about the model parameters. However, with survey data a theoretically derived model rarely holds for all observations. First, the model may not be appropriate for every subgroup in the population; second, some relevant explanatory variables may not be measured in the survey; third, the true relations among the variables may not be exactly linear. In addition, informative nonresponse can distort the model relationship 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 estimators is the ability to say we are estimating a population quantity with the price of generally larger estimated variances than for OLS. If the working model is good, we expect that the point estimators $\hat{\boldsymbol{\beta}}$ and $\mathbf{b}$ should be similar. However, if the model is misspecified, survey-weighted and OLS estimates can be far apart as illustrated in Korn and Graubard (1995). In this dissertation, I assume that analysts will use survey weights to estimate regression models. The diagnostics to be developed account for the effects of these weights.

Chapter 3: Identification of Single Influential Observations

3.1 Introduction

Examples from real surveys show that there is a need for influence diagnostics since a small number of the sampled units with possible extreme values could play a crucial role in the estimation of statistics and their variances. In 1986, the Joint Economic Committee of the U.S. Congress released a study indicating a sharp increase in the percentage of wealth held by the most affluent families in America. The richest 0.5% of families was estimated to hold 35% of the wealth in 1983, whereas in 1963 this proportion was 25%. The finding was proved to be wrong because a respondent with a very large weight was recorded to have $200 million in wealth attributed to him when the correct number was $2 million (Ericksen, 1988). The estimated share of wealth by the richest 0.5% of families dropped to 27% after the figure was corrected.

As in other statistical disciplines, outliers have been a well-known problem in design-based survey sampling (Lee, 1995). Usually outliers feature extreme values that may be substantially different from the bulk of the data. Chambers (1986) characterized outliers into two basic types: nonrepresentative and representative. The former means the value for a sample unit is incorrect or the value is unique to a particular population unit, whereas the latter refers to cases in which the values are correct and there are others like them in the nonsample part of the population. Sometimes the reported observations in sample surveys are named as influential because inclusion or exclusion of them can greatly change the parameter estimates. There are diverse reasons for survey data containing influential observations, such as editing error, observation error, or simply a skewed population. T. M. F. Smith (1987) pointed out, “individual values can be influential in randomization inference either when they are included in the sample or when they are not in the sample,” and “diagnostics are useful in the former case.” A few nonsample, nonrepresentative outliers, for example, can have a large effect on the error of an estimated total but cannot be identified by diagnostics. Extreme values and

influential values may not necessarily refer to the same observations due to sizes of sample weights. The distinction of the two concepts has been noted by some survey researchers (see Gambino 1987, Srinath 1987, and Bruce 1991). The premise in this research is that an analyst will be looking for a linear regression model that fits reasonably well for the bulk of the finite population. We have in mind two general goals. First, the influence diagnostics should allow the analyst to identify points that may not follow that model and have an influence on the size of estimated model parameters, or their estimated standard errors, or both. Second, the diagnostics should identify points that are influential in PML estimation because of the way the sample was selected; in particular, because of the size of the survey weights. These two goals sometimes conflict. For example, a point that is influential in the population may not be influential in the sample if its weight is small. The reverse is also true.

3.2 Basic Idea in Influence Assessment

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. They mention three questions in designing methods for influence analysis: the perturbation scheme, the particular aspect of an analysis to monitor, and the method of measurement. Different answers to these questions can lead to a variety of different diagnostics. For example, if we consider only one perturbation scheme in which the data are modified by deletion of cases and we want to monitor how the deletion will affect the estimation of regression coefficients, we may formulate relevant statistics to measure the effect of deletion.

Conventional model-based influence diagnostics mainly use the technique of row deletion, determining if the fitted regression function is dramatically changed when one or multiple observations are discarded. The statistics which are widely adopted include DFBETAS, DFFITS, Cook's Distance, COVRATIO, and so on (e.g. see Neter, Kutner, Nachtsheim, and Wasserman 1996).

These statistics do not have immediate application to randomization inference for sample surveys. As Brewer and Särndal (1983) noted, the idea of robustness to

departures from an assumed model does not fit naturally into a purely design-based framework, because models are not used directly in inference. However, the consideration of a model is needed to motivate the use of diagnostic statistics in finite population inference. The goal of inference will be to develop procedures that permit "good" estimates of parameters for a model that fits reasonably well for most of a finite population. By omitting influential points, ideally, a less design-biased and more stable estimates of underlying model parameters will result. Even in the prediction approach, the inclusion of sampling weights and the application of robust variance estimation mean that standard diagnostics need adaptation.

3.3 Sources of Influence in Survey Data

The influence of observations on regression estimation under the survey setting may come from at least three sources: outlying $\mathbf{Y}$ values, $\mathbf{X}$ values, and sampling weights $\mathbf{W}$. Atypical or extreme values of any of these or 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:

1. 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.
2. 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.
3. Given the large sample size in many surveys it would be important to set up some criteria to single out the influential units, or groups of units, instead of only reporting diagnostics for all units in the sample.

A natural question is how large a particular measure of influence should be so that an observation should receive special treatment. Belsley, Kuh and Welsch (1980) recommended choosing reasonable cutoffs by judgment and intuition, combining

empirical and theoretical arguments. Under the survey setting, we may not be able to directly borrow the cutoffs for the conventional regression diagnostic statistics if they are not carefully justified. New methods of determining the cutoffs need to be adapted to complex survey designs.

3.4 Review of Traditional Techniques

The conventional diagnostic techniques are developed to examine whether a given dataset is in accordance with the conditions of regression model (2.1).

3.4.1 Leverages and Residuals

In the conventional model diagnostics, the residuals, $\mathbf{e} = \mathbf{Y} - \mathbf{Xb}$, and the hat matrix, $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$, are the measures used to identify the outlying $\mathbf{Y}$ and $\mathbf{X}$ values, respectively. The diagonal element $h_{ii} = \mathbf{x}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i$ of the hat matrix is called the leverage of the *i*th case which is the weight of observation Y_i in determining the fitted value $\hat{Y}_i$. It has the following properties: (1) $0 \le h_{ii} \le 1$; (2) $\sum_{i=1}^{n} h_{ii} = p$, where p is the number of columns in $\mathbf{X}$ matrix. A leverage value h_{ii} is usually considered as large if it is more than twice their mean, $\bar{h} = \frac{p}{n}$. The residuals are often rescaled relative to their standard errors. The ratio of e_i to $s(e_i) = \sqrt{s^2(1-h_{ii})}$, where $s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$ is the mean square error, is called the internally studentized residual and denoted by r_i. Replacing s^2 with $s^2(i)$, the mean square error when the *i*th case is omitted in fitting the regression function, we obtain an externally studentized residual $r_i^* = \frac{e_i}{s(i)\sqrt{1-h_{ii}}}$ which follows the *t* distribution with $n-p-1$ degrees of freedom assuming that model

(2.1) holds, including the assumption of normal errors.

3.4.2 Influence on Regression Coefficients: DFBETA and DFBETAS

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

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

where $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}$ is the (*jj*)th element of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. The denominator of $DFBETAS_{ij}$ is analogous to the estimated standard error of $\mathbf{b}$ with the sample standard error s replaced by the delete-one version $s(i)$. The DFBETAS statistic is the product of a quantity of order $n^{-1/2}$, a *t* distributed random variable, and a quantity that approaches 1 (assuming $h_{ii} \to 0$). Belsley, Kuh and Welsch (1980) propose a cutoff point of $\frac{2}{\sqrt{n}}$ to identify influential cases. Thus, if all the observations in the sample follow an underlying normal model, the $\mathbf{X}$'s are bounded, and the leverages are small, roughly 95% of the observations will have a DFBETAS statistic less than $\frac{2}{\sqrt{n}}$ in absolute value. In some samples, especially small or moderate size ones, this statement is less precise since h_{ii} may not be negligible and the term involving c_{ji} may not be near $\frac{1}{\sqrt{n}}$.

DFBETAS is somewhat cumbersome to work with because an analyst must examine *pn* values. For each observation *i*, there are *p* DFBETAS – one for each parameter.

3.4.3 Influence on Fitted Values: DFFIT and DFFITS

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

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

A large value of DFFITS indicates that the observation is very influential in its neighborhood of the $\mathbf{X}$ space. A general cutoff to consider is 2; a size-adjusted cutoff recommended by Belsley, Kuh, and Welsch (1980) is $2\sqrt{\frac{p}{n}}$, where $\frac{p}{n}$ is the mean leverage.

3.4.4 Cook's Distance

Cook's distance provides an overall measure of the combined impact of an observation on all of the estimated regression coefficients $\mathbf{b}$ (e.g. see Cook 1977 and Weisberg 1985). It can be derived from the confidence region of $\boldsymbol{\beta}$, which at level $100(1-\alpha)\%$ is given by those values $\mathbf{b}^*$ satisfying

$$\frac{\left(\mathbf{b}^* - \mathbf{b}\right)^T \mathbf{X}^T\mathbf{X}\left(\mathbf{b}^* - \mathbf{b}\right)}{ps^2} \le F\left(1-\alpha; p, n-p\right).$$

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

$$D_i = \frac{\left(\mathbf{b}(i) - \mathbf{b}\right)^T \mathbf{X}^T\mathbf{X}\left(\mathbf{b}(i) - \mathbf{b}\right)}{ps^2}.$$

This is a measure of the distance from $\mathbf{b}(i)$ to $\mathbf{b}$. If $\mathbf{b}(i)$ and $\mathbf{b}$ are relatively far from each other, this means that unit i has a substantial effect on the full sample estimate. Large values of D_i indicate observations that are influential on joint inferences about all the parameters in the linear model. It has been found useful to relate D_i to the percentile values of $F(1-\alpha; p, n-p)$ distribution to make the judgment on influence. For example, if the percentile value is less than about 10 or 20 percent, the unit has little apparent influence on the regression coefficients. On the other hand, if the percentile value is near 50 percent or more, the influence is potentially important.

A more convenient form for D_i, without fitting a new regression function for each deletion, follows from substitution for DFBETA and yields

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

where $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$ is the internally studentized residual. Note from this expression that D_i depends on the size of the studentized residual and the leverage value. Atkinson (1982) suggested replacing s^2 by the deletion estimate $s^2(i)$, scaling the statistic by the average leverage $\frac{p}{n}$, and then taking the square root to give a residual like quantity. The resulting modified Cook statistic is

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

where r_i^* is the externally studentized residual. It can be derived that, if n is extremely large, the cutoff of the modified Cook's distance is 2 because r_i^* is t distributed.

3.5 Variance Estimation Methods for Complex Survey Data

An important issue in influence analysis is the cutoff value to be used in determining what points are influential. In the case of OLS estimation, we have seen that some diagnostic statistics are formulated using variance estimates of $\hat{\boldsymbol{\beta}}$ and cutoff points are developed in terms of some distributions. For example, the standard diagnostic DFBETA is scaled by dividing by an estimate of the model standard error of $\hat{\boldsymbol{\beta}}$. When the sample is associated with survey design features such as stratification, clustering, and other complexities, there are choices on how to construct diagnostic statistics. We propose three options here, using DFBETAS as an illustration:

(i) Ignore all design complexities and use the OLS construction to estimate both $\boldsymbol{\beta}$ and DFBETAS. This would be defensible if strictly model-based analysis were being done and the underlying model were (2.1). The design could be at least partially accounted for by incorporating design variables like stratum indicators in $\mathbf{x}_i$.

(ii) Estimate $\boldsymbol{\beta}$ using the Survey Weighted estimator. Standardize DFBETA by dividing by a standard error that would be appropriate to estimate the design-based standard error of $\hat{\boldsymbol{\beta}}$ if the sample had been selected with varying probabilities and with replacement in singe-stage, unstratified, unclustered sampling. Depending on how it is constructed, this type of variance estimator can be appropriate for a certain class of models.

(iii) Estimate $\boldsymbol{\beta}$ using the Survey Weighted estimator. Standardize DFBETA by dividing by a standard error that would be appropriate to estimate the design-based standard error of $\hat{\boldsymbol{\beta}}$, approximately accounting for stratification, clustering, and unequal weighting. As in (ii), this type of variance estimator can also be appropriate for a certain class of models.

Variance estimates for options (ii) and (iii) are discussed in more details below. The following notations will be used throughout this Chapter:

s : cluster sample, or unit sample for single-stage sampling;

s_i : unit sample in cluster *i*;

U : universe of clusters;

U_i : universe of units in cluster *i*;

n : number of sample clusters, or number of sample units for single-stage sampling;

N : number of clusters in universe, or number of units in universe for single-stage sampling;

m_i : number of sample units in cluster *i*;

M_i : number of population units in cluster *i*;

$h = 1,...,H$: index of strata. Subscript h denotes the statistics for stratum h;

$i, i' = 1,...,n$: index of sample clusters, or sample units for single-stage sampling;

$k, k' = 1,...,m_i$: index of sample units in cluster *i*.

Hence, w_{hik}, $\mathbf{x}_{hik}$, and Y_{hik}, respectively, are the sample weight, the vector of auxiliary variables, and the value of the dependent variable for the kth unit within cluster i of stratum h; m_{hi} and M_{hi} are the number of units in cluster *i* of stratum h in the sample and in the population;

3.5.1 Asymptotic Framework

In order to develop the distributional properties of the statistics such as DFBETAS, DFFITS, and so on, we need some assumptions for orders of magnitudes. An asymptotic framework needs to be specified since, although the population in a survey problem may be very large, it is still finite (Shao, 1996). We assume that the finite population under study is a member of a sequence of finite populations indexed by

$t = 1, 2, \ldots$, but t will be suppressed in order to simplify the notation. The total number of first-stage sampled clusters or primary sampling units (PSUs), n, is assumed large, that is, $n = \sum_h n_h \to \infty$ as $t \to \infty$, where n_h is the number of sampled clusters within stratum *h*. This includes two common situations in surveys: first, all the n_h are small (or bounded) but H is large, e.g., an extreme case is the design of two PSUs per stratum; second, all the n_h are large but H is bounded. We assume that no survey weight is disproportionately large, or

$$\max_{h,i,k} \frac{m_{hi} w_{hik} n}{N} = O(1) \tag{3.1}$$

where m_{hi} is the number of sampled units in the *i*th cluster of stratum *h*. If the sampling design is stratified two-stage sampling and simple random sampling is used in both stages of sampling, then $w_{hik} = \frac{N_h M_{hi}}{n_h m_{hi}}$ and (3.1) reduces to

$$\max_{h,i} \frac{N_h M_{hi} n}{n_h N} = O(1) \tag{3.2}$$

where M_{hi} is the number of units in the *i*th cluster of stratum *h* in universe, and N_h is the number of clusters within stratum *h* in universe. Condition (3.2) becomes $\max_{h,i} \frac{N_h n}{n_h N} = O(1)$ if M_{hi} is bounded. The two common situations in surveys mentioned above satisfy this assumption about the survey weights. More specifically, using (3.2) as an example, the condition will be satisfied if as $t \to \infty$,

Case 1: $H \to \infty$, $\frac{N_h}{n_h}$ is bounded, $\frac{N}{H}$ and $\frac{n}{H}$ converge to constants.

Case 2: H is bounded, $\frac{n_h}{n}$ and $\frac{N_h}{N}$ converge to positive constants.

Based on the above assumptions and, again assuming the $\mathbf{X}$'s are bounded, we can derive the following orders of magnitude for several aggregate quantities:

(1) $\mathbf{A} = \mathbf{X}^T \mathbf{W} \mathbf{X} = \sum_h \sum_{i \in s_h} \sum_{k \in s_{hi}} \mathbf{x}_{hik} w_{hik} \mathbf{x}_{hik}^T = O(N)$, and $\mathbf{A}^{-1} = O\left(N^{-1}\right)$, elementwise;

(2) $\mathbf{C} = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{W} = O\left(n^{-1}\right)$, elementwise;

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

We take the equal signs in expressions (2) and (3) when m_{hi} are bounded.

For single-stage sampling, assumption (3.1) reduces to $\max\left(w_i n/N\right) = O(1)$, where *n* and *N* are sample size and population size, respectively. The three conditions above still hold. Note that it may be possible to relax the assumption that $\mathbf{X}$ is bounded (e.g., see Miller 1989) and still obtain (1)-(3) above.

3.5.2 Variance Estimation for Single-Stage Sampling With Replacement

Assume the working model is (2.1). Treating the finite population as a sample of size *N* from 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 $e_{iU} = Y_i - \mathbf{x}_i^T\mathbf{B}$. Let E_M denote an expectation with respect to model (2.1). Since

$$E_M\left(e_{iU}^2\right) = \sigma^2\left(1 - h_{ii}^U\right), \text{ where } h_{ii}^U = \mathbf{x}_i^T\left(\mathbf{X}_N^T\mathbf{X}_N\right)^{-1}\mathbf{x}_i,$$

$$\text{and } E_M\left(\sum_{i\in U} e_{iU}^2\right) = \sigma^2\left(N - \sum_{i\in U} h_{ii}^U\right) = \sigma^2\left(N-p\right),$$

σ_U^2 is an unbiased estimate of σ^2 with respect to the working model. 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 an estimator $\hat{\sigma}^2 = \frac{1}{\hat{N}}\sum_{i\in s} w_i e_i^2$, where e_i is the sample residual defined as $e_i = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, and $\hat{N} = \sum_{i\in s} w_i$. The statistic $\hat{\sigma}^2$ is an approximately design unbiased estimator for σ_U^2 and, if the working model is correctly specified, is also estimating σ^2. In the following we sketch the reason for the approximate unbiasedness and suggest using a modified version,

$$\hat{\sigma}^2 = \frac{1}{\hat{N} - p} \sum_{i \in s} w_i e_i^2 \ . \tag{3.3}$$

We have $\hat{\boldsymbol{\beta}} = \mathbf{B} + \mathbf{C}$ (Fuller 2002, Fuller and Isaki 1982) where $\mathbf{C} = O_p\left(1/\sqrt{n}\right)$ elementwise, $N/\hat{N} = 1 + O_p\left(1/\sqrt{n}\right)$, and let E_π denote expectation with respect to the sample design. Then, we have

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{N}{\hat{N}} \frac{1}{N} \sum_{i \in s} w_i e_i^2 \\
&= \left(\frac{N}{\hat{N}}\right) \frac{1}{N} \sum_{i \in s} w_i \left[\left(Y_i - \mathbf{x}_i^T \mathbf{B}\right)^2 - 2\mathbf{x}_i^T \mathbf{C}\left(Y_i - \mathbf{x}_i^T \mathbf{B}\right) + \left(\mathbf{x}_i^T \mathbf{C}\right)^2 \right] \\
&= \left(1 + \frac{N - \hat{N}}{\hat{N}}\right) \left[\frac{1}{N} \sum_{i \in s} w_i \left(Y_i - \mathbf{x}_i^T \mathbf{B}\right)^2 - \frac{2}{N} \sum_{i \in s} w_i \mathbf{x}_i^T \mathbf{C}\left(Y_i - \mathbf{x}_i^T \mathbf{B}\right) + \frac{1}{N} \sum_{i \in s} w_i \left(\mathbf{x}_i^T \mathbf{C}\right)^2 \right] \\
&= \frac{1}{N} \sum_{i \in s} w_i \left(Y_i - \mathbf{x}_i^T \mathbf{B}\right)^2 + O_p\left(1/\sqrt{n}\right).
\end{aligned}$$

Under some technical conditions, the expectation of the $O_p\left(1/\sqrt{n}\right)$ term is itself $O\left(1/\sqrt{n}\right)$, e.g. if the $O_p\left(1/\sqrt{n}\right)$ term is uniformly integrable (see Serfling 1980, Thm. C, p.15). Consequently,

$$E_\pi\left(\hat{\sigma}^2\right) \doteq \frac{1}{N} \sum_{i \in U} \left(Y_i - \mathbf{x}_i^T \mathbf{B}\right)^2 = \frac{N - p}{N} \sigma_U^2 \doteq \sigma_U^2,$$

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

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 under (2.1),

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

can be estimated as

$$v_M\left(\hat{\boldsymbol{\beta}}\right)=\hat{\sigma}^2\mathbf{A}^{-1}\left(\sum_{i=1}^{n}w_i^2\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{A}^{-1}. \tag{3.4}$$

If model (2.1) is misspecified, instead let us consider a model in which the Y_i's are independent but whose variances differ among the units:

$$Y_i=\mathbf{x}_i^T\boldsymbol{\beta}+\varepsilon_i,\ \varepsilon_i\sim ind(0,\psi_i), \tag{3.5}$$

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

$$V_M\left(\hat{\boldsymbol{\beta}}\right)=\mathbf{A}^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i w_i\psi_i w_i\mathbf{x}_i^T\right)\mathbf{A}^{-1}. \tag{3.6}$$

The associated residual for unit i is $e_i=Y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. Under model (3.5), the squared residual has the expectation

$$E_M\left(e_i^2\right)=\psi_i\left(1-h_{ii}\right)^2+\sum_{i'\neq i}h_{ii'}^2\psi_{i'}$$

with $h_{ii'}$ being the (ii')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 (3.6) by the squares of the corresponding residuals e_i^2 based on the regression fit, the sandwich estimator of the unknown model variance is

$$v_W\left(\hat{\boldsymbol{\beta}}\right)=\mathbf{A}^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i w_i e_i^2 w_i\mathbf{x}_i^T\right)\mathbf{A}^{-1}=\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}diag(e_i^2)\mathbf{W}\mathbf{X}\mathbf{A}^{-1}. \tag{3.7}$$

Using $\mathbf{C}=\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}=\left(c_{ji}\right)_{p\times n}$ as defined in Section 3.5.1, we have $v_W(\hat{\boldsymbol{\beta}}_j)=\sum_{i=1}^{n}c_{ji}^2e_i^2$.

This estimator is model robust against deviations from the constant variance structure as in model (2.1). 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

$$\hat{\boldsymbol{\beta}} - \mathbf{B} \doteq \mathbf{A}_N^{-1} \sum_{i \in s} \mathbf{x}_i w_i \left(Y_i - \mathbf{x}_i^T \mathbf{B} \right) = \sum_{i \in s} \mathbf{z}_i \tag{3.8}$$

where $\mathbf{B}$ is the finite population regression parameter, $\mathbf{A}_N = \mathbf{X}_N^T \mathbf{X}_N$, 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 can be approximated by single-stage with-replacement sampling, the linear substitute approach can be used to obtain the design consistent variance estimator

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

where $\mathbf{z}_i^* = \mathbf{A}^{-1} \mathbf{x}_i w_i e_i$ and $\bar{\mathbf{z}}^* = \frac{1}{n} \sum_s \mathbf{z}_i^*$ (e.g., see the SUDAAN v.8 manual). Like v_W, v_L is model-robust since it is approximately unbiased under the general model (3.5). Next, note that

$$\bar{\mathbf{z}}^* = \frac{1}{n} \sum_s \mathbf{A}^{-1} \mathbf{x}_i w_i e_i = \frac{1}{n} \sum_s \mathbf{A}^{-1} \mathbf{x}_i w_i \left(Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right) = \frac{1}{n} \sum_s \left(\mathbf{A}^{-1} \mathbf{x}_i w_i Y_i - \mathbf{A}^{-1} \mathbf{x}_i w_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right) = \mathbf{0}$$

where we use the fact that $\mathbf{A} = \sum_s \mathbf{x}_i w_i \mathbf{x}_i^T$. Then $v_L \left(\hat{\boldsymbol{\beta}} \right) = \frac{n}{n-1} \sum_{i=1}^{n} \mathbf{z}_i^* \mathbf{z}_i^{*T}$, implying that v_L and v_W are approximately the same when the sample size is large enough that $\frac{n}{n-1} \approx 1$.

If the design uses stratification, the notation above needs elaboration. Let $\mathbf{x}_{hi}$ be the vector of independent variables for unit i in stratum h, w_{hi} be the weight for that unit, and $e_{hi} = Y_{hi} - \mathbf{x}_{hi}^T \hat{\boldsymbol{\beta}}$. If stratum dummies are not part of the model, then

$$v_L \left(\hat{\boldsymbol{\beta}} \right) = \sum_h \frac{n_h}{n_h - 1} \sum_{i \in s_h} \left(\mathbf{z}_{hi}^* - \bar{\mathbf{z}}_h^* \right) \left(\mathbf{z}_{hi}^* - \bar{\mathbf{z}}_h^* \right)^T$$

which uses $\mathbf{z}_{hi}^* = \mathbf{A}^{-1} \mathbf{x}_{hi} w_{hi} e_{hi}$ and $\bar{\mathbf{z}}_h^* = \sum_{i \in s_h} \mathbf{z}_{hi}^* \Big/ n_h$. In that case, $\bar{\mathbf{z}}_h^*$ is not $\mathbf{0}$, but

$E_M\left(\bar{\mathbf{z}}_h^*\right)=\mathbf{0}$. The comparison of v_L and v_W in this case is discussed in more detail in Section 3.5.3 in the context of stratified cluster sampling.

3.5.3 Variance Estimation for Multistage Sampling Design

For a multistage area probability sample, the design variance will be computed to account for the complexity of the design assuming the first-stage sample was selected with replacement. The analogous model-based assumption is that units in different PSUs are independent under a model. Suppose there are $i=1,\ldots,N$ clusters in the population and $k=1,\ldots,M_i$ units in cluster *i*. Note that clustered samples often use multiple stages of selection, but users are typically provided only identifiers for one type of cluster. As a result, considering only one level of clustering will match the level of detail available to most users. Suppose that $\mathbf{x}_{ik}$ is a *p*-vector of explanatory variables for unit *k* in cluster *i*. The linear model is

$$\begin{aligned} &Y_{ik}=\mathbf{x}_{ik}^T\boldsymbol{\beta}+\varepsilon_{ik} \qquad i=1,\ldots,N,\quad k=1,\ldots,M_i,\\ &Cov_M\left(\varepsilon_{ik},\varepsilon_{i'k'}\right)=\begin{cases}\sigma^2 & i=i',k=k'\\ \sigma^2\rho & i=i',k\neq k'\\ 0 & i\neq i',k\neq k'\end{cases}. \end{aligned} \tag{3.9}$$

This model posits that all units have a common variance and the intracluster correlation, ρ, is the same for all clusters. Units in different clusters are uncorrelated. In principle, ρ in (3.9) can be negative and has a lower bound of $-(D-1)^{-1}$ where D is defined below (Valliant, Dorfman, and Royall, 2000). In practice, ρ is usually positive and can be estimated using analysis of variance methods, as described in Section 5.3.2.

In order to compute standardized residuals, we will need estimates of the parameters in (3.9). This model is restrictive but is used only to get cutoff values for diagnostic statistics in Section 3.6.2, 3.6.3, and 3.6.4. For other analyses, we can use variance estimators for $\hat{\boldsymbol{\beta}}$ that do not depend on such a restrictive model.

In the clustered case, 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*) 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*, $i = 1, \ldots, n$

$\mathbf{W}_i$ = the $m_i \times m_i$ diagonal matrix of survey weights, and

$\mathbf{Y}_i$ = the m_i-vector of Y_{ik}'s.

Using these definitions, $\mathbf{A}$ can also be written as $\mathbf{A} = \sum_{i \in s} \mathbf{X}_i^T \mathbf{W}_i \mathbf{X}_i$.

If we treat the finite population as a sample, under model (3.9), the variance parameters are estimated as

$$\left[(1-\rho)\sigma^2 \right]_U = P$$
$$\left[\rho\sigma^2 \right]_U = (Q - P)/D$$
$$D = \left(M_+ - \sum_{i \in U} M_i^2 / M_+ \right) \Big/ (N-1)$$

where

$$P = \frac{1}{N} \sum_{i \in U} \frac{1}{M_i - 1} \sum_{k \in U_i} \left(e_{ik}^U - \bar{e}_i^U \right)^2$$

$$Q = \frac{1}{N-1} \sum_{i \in U} M_i \left(\bar{e}_i^U - \bar{e}^U \right)^2$$

$$M_+ = \sum_i M_i$$

$$e_{ik}^U = Y_{ik} - \mathbf{x}_{ik}^T \mathbf{B}, \quad \bar{e}_i^U = \sum_{k \in U_i} e_{ik}^U / M_i, \quad \bar{e}^U = \sum_i \bar{e}_i^U M_i / M_+ .$$

The notation $[\bullet]_U$ means that the quantity in the brackets is a finite population parameter. We have $E_M(P) \doteq (1-\rho)\sigma^2$ and $E_M(Q) \doteq (1-\rho)\sigma^2 + D \cdot \rho\sigma^2$. Here

we assume that e_{ik}^{U} is a good estimate of ε_{ik} in model (3.9). Proofs refer to Valliant et al. (2000), p258.

As in the single-stage sampling case, our goal is to find the design-based estimates of $\left[(1-\rho)\sigma^2\right]_U$ and $\left[\rho\sigma^2\right]_U$, or P, Q, and D for the two-stage sampling design. Pfeffermann et al. (1998) proposed the probability-weighted iterative generalized least squares (PWIGLS) estimator to obtain consistent estimates of the variance parameters $\left[(1-\rho)\sigma^2\right]_U$ and $\left[\rho\sigma^2\right]_U$ from the two-level model. The PWIGLS estimator, which assumes that the sampling probabilities for both stages π_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

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

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

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

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

estimators of $\left[(1-\rho)\sigma^2\right]_U$ and $\left[\rho\sigma^2\right]_U$, respectively, as

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

Another alternative is to use analysis of variance or restricted maximum likelihood methods. An application of this using SAS PROC VARCOMP (PROC MIXED can also be used) is discussed later in Section 5.3.2.

When $\widehat{\rho\sigma^2}$ and $\widehat{(1-\rho)\sigma^2}$ (or $\hat{P}$, $\hat{Q}$, and $\hat{D}$) are available, the estimated variance of $\hat{\boldsymbol{\beta}}$ under (3.9) can be constructed as

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

It follows that an estimate of ρ is

$$\hat{\rho} = \frac{\widehat{\rho\sigma^2}}{\widehat{(1-\rho)\sigma^2} + \widehat{\rho\sigma^2}} \quad \text{or} \quad \hat{\rho} = \left[\frac{\hat{P}\hat{D}}{\hat{Q}-\hat{P}}+1\right]^{-1}. \tag{3.10}$$

This estimator is highly dependent on the working model and is not robust to departures from that model. Note that $\hat{\rho}$ is not necessarily confined to $\left[-(D-1)^{-1},1\right]$ when analysis of variance methods are used to estimate $\rho\sigma^2$ and $(1-\rho)\sigma^2$. If $\hat{\rho}$ in (3.10) is outside $\left[-(D-1)^{-1},1\right]$, the usual procedure is to assign it the nearest boundary value.

As in the case of estimation under the single-stage sampling model, we can construct a simple sandwich estimator that is consistent under a reasonably general variance specification. Consider the model:

$$\begin{aligned}&E_M\left(Y_{ik}\right) = \mathbf{x}_{ik}^T\boldsymbol{\beta} && i=1,\ldots,N,\ \ k=1,\ldots,M_i \\ &Cov_M\left(Y_{ik},Y_{i'k'}\right) = 0 && i\neq i'\end{aligned}. \tag{3.11}$$

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 $e_{ik} = Y_{ik} - \mathbf{x}_{ik}^T \hat{\boldsymbol{\beta}}$. Define the hat matrix as

$$\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W} = \begin{bmatrix} \mathbf{X}_1\mathbf{A}^{-1}\mathbf{X}_1^T\mathbf{W}_1 & \cdots & \mathbf{X}_1\mathbf{A}^{-1}\mathbf{X}_n^T\mathbf{W}_n \\ \vdots & & \vdots \\ \mathbf{X}_n\mathbf{A}^{-1}\mathbf{X}_1^T\mathbf{W}_1 & \cdots & \mathbf{X}_n\mathbf{A}^{-1}\mathbf{X}_n^T\mathbf{W}_n \end{bmatrix}$$

and let $\mathbf{H}_{ii'} = \mathbf{X}_i\mathbf{A}^{-1}\mathbf{X}_{i'}^T\mathbf{W}_{i'}$. Then the vector of residuals for sample cluster *i* is $\mathbf{e}_i = \mathbf{Y}_i - \sum_{i' \in s} \mathbf{H}_{ii'}\mathbf{Y}_{i'} = \left(\mathbf{I}_{m_i} - \mathbf{H}_{ii}\right)\mathbf{Y}_i - \sum_{i' \neq i} \mathbf{H}_{ii'}\mathbf{Y}_{i'}$. We have

$$E_M\left(\mathbf{e}_i\mathbf{e}_i^T\right) = \left(\mathbf{I}_{m_i} - \mathbf{H}_{ii}\right)V_M\left(\mathbf{Y}_i\right)\left(\mathbf{I}_{m_i} - \mathbf{H}_{ii}\right)^T + \sum_{i' \neq i} \mathbf{H}_{ii'}V_M\left(\mathbf{Y}_i\right)\mathbf{H}_{ii'}^T. \quad (3.12)$$

If $\mathbf{A}^{-1} = O\left(N^{-1}\right)$, and the sample sizes m_i are bounded, then $\mathbf{H}_{ii'} = O\left(n^{-1}\right)$. Thus, as the number of sampled PSUs becomes large, or $n \to \infty$, $E_M\left(\mathbf{e}_i\mathbf{e}_i^T\right) \cong V_M\left(\mathbf{Y}_i\right)$, and, consequently, the sandwich variance estimator is

$$v_W\left(\hat{\boldsymbol{\beta}}\right) = \sum_{i \in s} \mathbf{A}^{-1}\mathbf{X}_i^T\mathbf{W}_i\left(\mathbf{e}_i\mathbf{e}_i^T\right)\mathbf{W}_i\mathbf{X}_i\mathbf{A}^{-1}. \quad (3.13)$$

Assuming the first-stage sample was selected with replacement, expression (3.8) becomes

$$\hat{\boldsymbol{\beta}} - \mathbf{B} \doteq \mathbf{A}_N^{-1}\sum_{i \in s}\sum_{k \in s_i} \mathbf{x}_{ik}w_{ik}(Y_{ik} - \mathbf{x}_{ik}^T\mathbf{B}) = \sum_{i \in s} \mathbf{z}_i, \quad (3.14)$$

where $\mathbf{z}_i = \mathbf{A}_N^{-1}\mathbf{X}_i^T\mathbf{W}_i(\mathbf{Y}_i - \mathbf{X}_i\mathbf{B})$. A design-based linearization estimator is given as

$$v_L\left(\hat{\boldsymbol{\beta}}\right) = \frac{n}{n-1}\left[\sum_{i=1}^n \mathbf{z}_i^*\mathbf{z}_i^{*T} - n\bar{\mathbf{z}}^*\bar{\mathbf{z}}^{*T}\right], \quad (3.15)$$

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

$$\bar{\mathbf{z}}^* = \frac{1}{n}\sum_s \mathbf{A}^{-1}\mathbf{X}_i^T\mathbf{W}_i\left(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right) = \frac{1}{n}\sum_s\left(\mathbf{A}^{-1}\mathbf{X}_i^T\mathbf{W}_i\mathbf{Y}_i - \mathbf{A}^{-1}\mathbf{X}_i^T\mathbf{W}\mathbf{X}_i\hat{\boldsymbol{\beta}}\right) = \mathbf{0}$$

using $\mathbf{A} = \sum_{i \in s} \mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i$. 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. We consider two cases here. First a simple model that reflects common intercepts and slopes among strata is

$$\begin{aligned} &E_M\left(Y_{hik}\right)=\mathbf{x}_{hik}^T\boldsymbol{\beta} \qquad h=1,...,H,\quad i=1,\ldots,N,\quad k=1,\ldots,M_{hi} \\ &Cov_M\left(Y_{hik},Y_{hi'k'}\right)=0 \qquad i\neq i' \end{aligned}. \quad (3.16)$$

Since clusters are assumed to be independently selected between and within strata, the regression estimator and its estimated variance would be similar to the ones derived from model (3.11) except that they include stratification and are expressed as sums over all clusters which are nested in strata. The survey weighted estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}=\sum_h\sum_{i\in s_h}\mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{Y}_{hi}$$

where the subscript hi refers to sample units in cluster *i*, stratum *h*, s_h is the set of sample clusters in stratum h, $\mathbf{X}_{hi}$ is the $m_{hi}\times p$ matrix of auxiliaries for sample cluster i in stratum h with m_{hi} being the number of sample units from cluster (hi). The components $\mathbf{W}_{hi}$ and $\mathbf{Y}_{hi}$ are defined by analogy to $\mathbf{W}_i$ and $\mathbf{Y}_i$ given earlier in this section. The model-based sandwich variance estimator is

$$v_W\left(\hat{\boldsymbol{\beta}}\right)=\sum_{h,i\in s_h}\mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T)\mathbf{W}_{hi}\mathbf{X}_{hi}\mathbf{A}^{-1}$$

where $\mathbf{e}_{hi}=\mathbf{Y}_{hi}-\mathbf{X}_{hi}\hat{\boldsymbol{\beta}}$.

After accounting for stratification, expression (3.14) becomes

$$\hat{\boldsymbol{\beta}}-\mathbf{B}\doteq\mathbf{A}_N^{-1}\sum_h\sum_{i\in s_h}\mathbf{X}_{hi}^T\mathbf{W}_{hi}(\mathbf{Y}_{hi}-\mathbf{X}_{hi}\mathbf{B})=\sum_h\sum_{i\in s_h}\mathbf{z}_{hi}$$

and the linearization variance estimator in this case is

$$\begin{aligned} v_L\left(\hat{\boldsymbol{\beta}}\right)&=\sum_h\frac{n_h}{n_h-1}\left[\sum_{i\in s_h}\left(\mathbf{z}_{hi}^*-\bar{\mathbf{z}}_h^*\right)\left(\mathbf{z}_{hi}^*-\bar{\mathbf{z}}_h^*\right)^T\right] \\ &=\sum_h\frac{n_h}{n_h-1}\left[\sum_{i\in s_h}\mathbf{z}_{hi}^*\mathbf{z}_{hi}^{*T}-n_h\bar{\mathbf{z}}_h^*\bar{\mathbf{z}}_h^{*T}\right] \end{aligned}$$

where n_h is the number of sample clusters in stratum h, $\mathbf{z}_{hi}^* = \mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{e}_{hi}$, and $\bar{\mathbf{z}}_h^* = \frac{1}{n_h}\sum_{i\in s_h}\mathbf{z}_{hi}^*$. This expression reduces to the formula for a single stage stratified design in Section 3.5.2 when the PSU sizes are all $n_{hi} = 1$. Like the sandwich estimator v_W, v_L is also approximately model unbiased for the variance of $\hat{\boldsymbol{\beta}}$, $V_M\left(\hat{\boldsymbol{\beta}}\right)$. The proof is illustrated as follows:

$$\mathbf{e}_{hi} = \left(\mathbf{I}_{m_{hi}} - \mathbf{H}_{hii}\right)\mathbf{Y}_{hi} - \sum_{i'\neq i}\mathbf{H}_{hii'}\mathbf{Y}_{hi'} \tag{3.17}$$

where $\mathbf{H}_{hii'} = \mathbf{X}_{hi}\mathbf{A}^{-1}\mathbf{X}_{hi'}^T\mathbf{W}_{hi'}$. Let $\boldsymbol{\Psi}_{hi} = V_M\left(\mathbf{Y}_{hi}\right)$, and assume $\mathbf{H}_{ii'} = O\left(n^{-1}\right)$, $\mathbf{W}_{hi} = O\left(\frac{N}{n}\right)$ and $\mathbf{A} = O\left(N\right)$. Then

$$E_M\left(\mathbf{e}_{hi}\mathbf{e}_{hi}^T\right) = \left(\mathbf{I}_{m_{hi}} - \mathbf{H}_{hii}\right)\boldsymbol{\Psi}_{hi}\left(\mathbf{I}_{m_{hi}} - \mathbf{H}_{hii}\right)^T + \sum_{i'\neq i}\mathbf{H}_{hii'}\boldsymbol{\Psi}_{hi'}\mathbf{H}_{hii'}^T = \boldsymbol{\Psi}_{hi} + O\left(n^{-1}\right).$$

Using (3.17) and $E_M\left(\mathbf{e}_{hi}\mathbf{e}_{hi'}^T\right) = Cov_M\left(\mathbf{e}_{hi}, \mathbf{e}_{hi'}^T\right)$, we have

$$\begin{aligned} E_M\left(\mathbf{e}_{hi}\mathbf{e}_{hi'}^T\right) &= -\left(\mathbf{I}_{m_{hi}} - \mathbf{H}_{hii}\right)\boldsymbol{\Psi}_{hi}\mathbf{H}_{hi'i}^T - \left(\mathbf{I}_{m_{hi'}} - \mathbf{H}_{hi'}\right)\boldsymbol{\Psi}_{hi'}\mathbf{H}_{hii'}^T + \sum_{i''\neq i,i'}\mathbf{H}_{hii''}\boldsymbol{\Psi}_{hi''}\mathbf{H}_{hi'i''}^T \\ &= O\left(n^{-1}\right). \end{aligned}$$

Then,

$$\begin{aligned} E_M\left(\mathbf{z}_{hi}^*\mathbf{z}_{hi}^{*T}\right) &= \mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}E_M\left(\mathbf{e}_{hi}\mathbf{e}_{hi}^T\right)\mathbf{W}_{hi}\mathbf{X}_{hi}\mathbf{A}^{-1} \\ &= \mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\left(\boldsymbol{\Psi}_{hi} + O\left(n^{-1}\right)\right)\mathbf{W}_{hi}\mathbf{X}_{hi}\mathbf{A}^{-1} = \mathbf{T}_{hi} + O\left(n^{-3}\right) = O\left(n^{-2}\right), \\ E_M\left(\mathbf{z}_{hi}^*\mathbf{z}_{hi'}^{*T}\right) &= \mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}E_M\left(\mathbf{e}_{hi}\mathbf{e}_{hi'}^T\right)\mathbf{W}_{hi'}\mathbf{X}_{hi'}\mathbf{A}^{-1} = O\left(n^{-3}\right) \end{aligned}$$

where $\mathbf{T}_{hi} = \mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\boldsymbol{\Psi}_{hi}\mathbf{W}_{hi}\mathbf{X}_{hi}\mathbf{A}^{-1}$. Hence we have

$$E_M\left(\bar{\mathbf{z}}_h^*\bar{\mathbf{z}}_h^{*T}\right)=\frac{1}{n_h^2}E_M\left(\sum_{i\in s_h}\mathbf{z}_{hi}^*\sum_{i\in s_h}\mathbf{z}_{hi}^{*T}\right)$$
$$=\frac{1}{n_h^2}\sum_{i\in s_h}E_M\left(\mathbf{z}_{hi}^*\mathbf{z}_{hi}^{*T}\right)+\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{i'\neq i}E_M\left(\mathbf{z}_{hi}^*\mathbf{z}_{hi'}^{*T}\right)$$
$$=\frac{1}{n_h^2}\sum_{i\in s_h}\mathbf{T}_{hi}+O\left(n^{-3}\right).$$

Let us consider the two cases mentioned in Section 3.5.1. If n_h is bounded,

$$E_M\left(\bar{\mathbf{z}}_h^*\bar{\mathbf{z}}_h^{*T}\right)\doteq\frac{1}{n_h^2}\sum_{i\in s_h}\mathbf{T}_{hi}\text{ , and}$$

$$E_M\left(v_L\right)=\sum_h\frac{n_h}{n_h-1}\left[E_M\left(\mathbf{z}_{hi}^*\mathbf{z}_{hi}^{*T}\right)-n_hE_M\left(\bar{\mathbf{z}}_h^*\bar{\mathbf{z}}_h^{*T}\right)\right]$$
$$\doteq\sum_h\frac{n_h}{n_h-1}\left[\sum_{i\in s_h}\mathbf{T}_{hi}-n_h\left(\frac{1}{n_h^2}\sum_{i\in s_h}\mathbf{T}_{hi}\right)\right]$$
$$=\sum_h\sum_{i\in s_h}\mathbf{T}_{hi}$$
$$=V_M\left(\hat{\boldsymbol{\beta}}\right).$$

The second term in brackets in $v_L\left(\hat{\boldsymbol{\beta}}\right)$ above (3.17) is $n_h\bar{\mathbf{z}}_h^*\bar{\mathbf{z}}_h^{*T}$. If $\frac{n_h}{n}$ converges to a constant, $n_hE_M\left(\bar{\mathbf{z}}_h^*\bar{\mathbf{z}}_h^{*T}\right)=O\left(n^{-2}\right)$ is negligible compared to $\sum_{i\in s_h}\mathbf{T}_{hi}$, and

$$E_M\left(v_L\right)\doteq\sum_h\frac{n_h}{n_h-1}\sum_{i\in s_h}\mathbf{T}_{hi}\doteq\sum_h\sum_{i\in s_h}\mathbf{T}_{hi}=Var_M\left(\hat{\boldsymbol{\beta}}\right).$$

Another way to account for stratification is to assume different linear models, or different slope parameters $\boldsymbol{\beta}_h$, in each stratum.

$$E_M\left(Y_{hik}\right)=\mathbf{x}_{hik}^T\boldsymbol{\beta}_h\qquad h=1,...,H\quad i=1,\ldots,N\quad k=1,\ldots,M_i.$$

Then, within each stratum, the estimation of regression parameters and their variances is the same as that for model (3.11).

$$\hat{\boldsymbol{\beta}}_h=\mathbf{A}_h^{-1}\sum_{i\in s_h}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{Y}_{hi}\text{ ,}$$

$$v_W\left(\hat{\boldsymbol{\beta}}_h\right)=\sum_{i\in s_h}\mathbf{A}_h^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\left(\mathbf{e}_{hi}\mathbf{e}_{hi}^T\right)\mathbf{W}_{hi}\mathbf{X}_{hi}\mathbf{A}_h^{-1}$$

where $\mathbf{A}_h=\sum_{i\in s_h}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{X}_{hi}$, and $\mathbf{e}_{hi}=\mathbf{Y}_{hi}-\mathbf{X}_{hi}\hat{\boldsymbol{\beta}}_h$. The design based linearization variance estimator of $\hat{\boldsymbol{\beta}}_h$ is similar to (3.15), but with a stratum subscript:

$$v_L\left(\hat{\boldsymbol{\beta}}_h\right)=\frac{n_h}{n_h-1}\sum_{i\in s_h}\left(\mathbf{z}_{hi}^*-\overline{\mathbf{z}}_h^*\right)\left(\mathbf{z}_{hi}^*-\overline{\mathbf{z}}_h^*\right)^T=\frac{n_h}{n_h-1}\sum_{i\in s_h}\mathbf{z}_{hi}^*\mathbf{z}_{hi}^{*T}\ .$$

When n_h is large v_W and v_L are approximately the same. The analysis of influence diagnostics will be conducted independently within each stratum for this setting.

3.6 Adaptations of Traditional Techniques to Regression on Complex Survey Data

In this section adapt the analysis of residuals, leverages, DFBETAS, DFFITS, and Cook's distance for use with complex survey data. Although some versions of these statistics are available in software that will fit weighted least squares regressions, the interpretation of them differs for survey data. Also, diagnostics that incorporate variance estimators must account for complex sample designs. Cutoff values must also be derived for the adapted statistics.

3.6.1 Residuals and Leverages

When survey weights are used in the regression, the predicted values become $\hat{\mathbf{Y}}=\mathbf{HY}$ and the residuals are $\mathbf{e}=\mathbf{Y}-\hat{\mathbf{Y}}=\left(\mathbf{I}-\mathbf{H}\right)\mathbf{Y}$, where the hat matrix includes the survey weights and, as in previous sections, is defined as

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

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, see

model (2.2), in the regression. However, results below can be modified to incorporate $\mathbf{V}$ simply by using $\mathbf{W}^* = \mathbf{W}\mathbf{V}^{-1}$ rather than $\mathbf{W}$. If we use the notations for single-stage sampling, the survey weighted hat matrix has the following properties:

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

2) $0 \le h_{ii} \le 1$;

3) $\sum_{i=1}^{n} h_{ii} = p$, where p is the number of columns in $\mathbf{X}$ matrix, and n is the total number of sample units;

(see Valliant et. al. (2000) for the proof of above properties.)

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

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

6) $\mathbf{HH} = \mathbf{H}$, and $\sum_{i'} h_{ii'}h_{i'i} = \sum_{i'} \mathbf{x}_i^T\mathbf{A}^{-1}\mathbf{x}_{i'}w_{i'}\mathbf{x}_{i'}^T\mathbf{A}^{-1}\mathbf{x}_i w_i = \mathbf{x}_i^T\mathbf{A}^{-1}\mathbf{x}_i w_i = h_{ii}$.

A large leverage may be caused by outlying $\mathbf{X}$ values, an outlying weight, or both. Similarly, a large residual may result from an outlying Y_i or w_i.

(1) Decomposition of Leverages

Leverages can be decomposed into components that separate the effect of the weight and the $\mathbf{X}$ values for a unit. We begin with a simple illustration. Suppose we have a simple model $y_i = \beta x_i + \varepsilon_i$, $\varepsilon_i \sim (0, \sigma^2 x_i)$. The WLS estimate of β is $b = \bar{y}_s / \bar{x}_s$ where we use $1/v_i = 1/x_i$ as the weight for this example. If we use superscript U to indicate the unweighted statistics, the WLS hat matrix is written as $\mathbf{H}^U = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{V}^{-1}$, where $\mathbf{A} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$, $\mathbf{V} = diag(x_i)_{n\times n}$, and $\mathbf{X} = (x_1, ..., x_n)^T$. The leverage of the ith observation is

$$\begin{aligned} h_{ii}^U &= x_i A^{-1} x_i v_i^{-1} \\ &= \frac{x_i}{n\bar{x}_s} x_i \frac{1}{x_i} \quad , \\ &= \frac{x_i}{n\bar{x}_s} \end{aligned}$$

since $\mathbf{A} = (x_1,...,x_n)\,diag\,(1/x_i)(x_1,...,x_n)^T = n\bar{x}_s$.

The parameter estimator accounting for survey weights is $\hat{\beta} = \mathbf{A}_W^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{Y}$, where $\mathbf{A}_W = \mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}$ and $\mathbf{W} = diag(w_i)_{n\times n}$. The $\mathbf{A}_W$ matrix can be simplified as follows,

$$\begin{aligned}\mathbf{A}_W &= (x_1,...,x_n)diag(w_i)diag(1/x_i)(x_1,...,x_n)^T \\ &= \sum_s w_i x_i \\ &= \hat{N}\bar{x}_W\end{aligned}$$

where $\hat{N} = \sum_s w_i = n\bar{w}$ and $\bar{x}_W = \sum_s w_i x_i / \sum_s w_i$. The weighted hat matrix is

$$\begin{aligned}\mathbf{H}^W &= \mathbf{X}\mathbf{A}_W^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1} \\ &= (x_1,...,x_n)^T\left(\hat{N}\bar{x}_W\right)^{-1}(x_1,...,x_n)\,diag\left(w_i/x_i\right) \\ &= \frac{1}{\hat{N}\bar{x}_W}\begin{pmatrix} w_1x_1 & w_2x_1 & \dots & w_nx_1 \\ & w_2x_2 & \dots & w_nx_2 \\ & & \ddots & \\ & & & w_nx_n \end{pmatrix}\end{aligned}$$

and the leverages on the diagonal of the hat matrix are defined as

$$h_{ii}^W = \frac{w_i x_i}{\hat{N}\bar{x}_W} = \frac{1}{n}\frac{w_i}{\bar{w}}\frac{x_i}{\bar{x}_W}.$$

Hence, the OLS leverage h_{ii}^U can be large if $x_i \gg \bar{x}_s$, whereas in the survey case the weighted leverage can be extreme if either $w_i \gg \bar{w}$ or $x_i \gg \bar{x}_W$.

Now let us extend the above analysis to a more general model (2.1):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \left(0, \sigma^2\mathbf{I}\right).$$

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 $\mathbf{x}_i^T = \left(x_{i1},\ldots,x_{i,p-1}\right)$ 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 $\mathbf{A}_W$ matrix is computed as

$$\mathbf{A}_W=\begin{pmatrix}\mathbf{1}^T\\ \mathbf{X}_1^T\end{pmatrix}\mathbf{W}\begin{pmatrix}\mathbf{1}\ \mathbf{X}_1\end{pmatrix}=\begin{pmatrix}\mathbf{1}^T\mathbf{W}\mathbf{1} & \mathbf{1}^T\mathbf{W}\mathbf{X}_1\\ \mathbf{X}_1^T\mathbf{W}\mathbf{1} & \mathbf{X}_1^T\mathbf{W}\mathbf{X}_1\end{pmatrix}\equiv\begin{pmatrix}\hat{N} & \hat{\mathbf{t}}_X^T\\ \hat{\mathbf{t}}_X & \mathbf{A}_{\mathrm{W1}}\end{pmatrix}$$

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 $\mathbf{A}_{\mathrm{W1}}$ is a $(p-1)\times(p-1)$ matrix. Using the inverse of a partitioned matrix,

$$\begin{aligned}\mathbf{A}_W^{-1}&=\begin{pmatrix}\frac{1}{\hat{N}}+\frac{1}{\hat{N}}\hat{\mathbf{t}}_X^T\mathbf{S}^{-1}\hat{\mathbf{t}}_X\frac{1}{\hat{N}} & -\frac{1}{\hat{N}}\hat{\mathbf{t}}_X^T\mathbf{S}^{-1}\\ -\frac{1}{\hat{N}}\mathbf{S}^{-1}\hat{\mathbf{t}}_X & \mathbf{S}^{-1}\end{pmatrix}\\ &=\begin{pmatrix}\frac{1}{\hat{N}}+\bar{\mathbf{x}}_W^T\mathbf{S}^{-1}\bar{\mathbf{x}}_W & -\bar{\mathbf{x}}_W^T\mathbf{S}^{-1}\\ -\mathbf{S}^{-1}\bar{\mathbf{x}}_W & \mathbf{S}^{-1}\end{pmatrix}\\ &=\begin{pmatrix}\frac{1}{\hat{N}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}+\begin{pmatrix}-\bar{\mathbf{x}}_W^T\\ \mathbf{I}\end{pmatrix}\mathbf{S}^{-1}\begin{pmatrix}-\bar{\mathbf{x}}_W & \mathbf{I}\end{pmatrix}\end{aligned}$$

where $\bar{\mathbf{x}}_W=\frac{\hat{\mathbf{t}}_X}{\hat{N}}$ is a $(p-1)\times 1$ vector and $\mathbf{S}=\mathbf{A}_{\mathrm{W1}}-\hat{\mathbf{t}}_X\hat{\mathbf{t}}_X^T\frac{1}{\hat{N}}$ is a $(p-1)\times(p-1)$ matrix. Simplifying the hat matrix using the above inverse matrix, we obtain

$$\begin{aligned}\mathbf{H}^W&=\mathbf{X}\mathbf{A}_W^{-1}\mathbf{X}^T\mathbf{W}\\ &=\begin{pmatrix}\mathbf{1} & \mathbf{X}_1\end{pmatrix}\left\{\begin{pmatrix}\frac{1}{\hat{N}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}+\begin{pmatrix}-\bar{\mathbf{x}}_W^T\\ \mathbf{I}\end{pmatrix}\mathbf{S}^{-1}\begin{pmatrix}-\bar{\mathbf{x}}_W & \mathbf{I}\end{pmatrix}\right\}\begin{pmatrix}\mathbf{1}^T\\ \mathbf{X}_1^T\end{pmatrix}\mathbf{W}\\ &=\left\{\frac{1}{\hat{N}}\mathbf{1}\mathbf{1}^T+\left(\mathbf{X}_1-\mathbf{1}\bar{\mathbf{x}}_W^T\right)\mathbf{S}^{-1}\left(-\bar{\mathbf{x}}_W\mathbf{1}^T+\mathbf{X}_1^T\right)\right\}\mathbf{W}\\ &=\left\{\frac{1}{\hat{N}}\mathbf{1}\mathbf{1}^T+\begin{pmatrix}\mathbf{x}_1^T-\bar{\mathbf{x}}_W^T\\ \vdots\\ \mathbf{x}_n^T-\bar{\mathbf{x}}_W^T\end{pmatrix}\mathbf{S}^{-1}\left(\mathbf{x}_1-\bar{\mathbf{x}}_W,\ldots,\mathbf{x}_n-\bar{\mathbf{x}}_W\right)\right\}\mathbf{W}.\end{aligned}$$

Then the leverage of *i*th observation, or the *i*th diagonal element of $\mathbf{H}^W$, is

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

Note that $\left(\mathbf{x}_i-\bar{\mathbf{x}}_W\right)^T\mathbf{S}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}_W\right)$ is an ellipsoid centered at $\bar{\mathbf{x}}_W$ (e.g. see Weisberg 1985), and $\hat{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}_W\right)^T\mathbf{S}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}_W\right)$ is the Mahalanobis distance from $\mathbf{x}_i$ to $\bar{\mathbf{x}}_W$. 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 X's, $\bar{\mathbf{x}}_W$.

If the error terms in the model have a general variance structure $\boldsymbol{\varepsilon}\sim\left(\mathbf{0},\sigma^2\mathbf{V}\right)$ with known $\mathbf{V}$ and unknown σ^2, the hat matrix is then defined as $\mathbf{H}^{WV}=\mathbf{X}\mathbf{A}_W^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}$ with

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

A formula for $\mathbf{A}_W^{-1}$ 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

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

In some applications, the individual values of some $\mathbf{X}$'s may be available for all units on the sample frame. This information could be helpful in deciding whether there are nonsample points with $\mathbf{X}$'s similar to high leverage points in the sample and in deciding whether such points should be removed when fitting the regression model.

(2) Residual Analysis

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 semi-studentized or studentized residual.

Assuming single-stage sampling, under model (2.1), the residual for unit i is $e_i = Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ and its model variance is $E_M\left(e_i^2\right) = \sigma^2\left[\left(1-h_{ii}\right)^2 + \sum_{i' \neq i} h_{ii'}^2\right]$. Since $h_{ii'} = O\left(n^{-1}\right)$, as we have demonstrated in Section 3.5.1, the term in the brackets has the form $1+o(1)$, and

$$E_M\left(e_i^2\right) \doteq \sigma^2. \tag{3.18}$$

Replacing σ^2 by its estimate $\hat{\sigma}^2 = \frac{1}{\hat{N}-p}\sum_{i \in s} w_i e_i^2$, we can standardize the residual for unit i as $\frac{e_i}{\hat{\sigma}}$ and compare it with a standard normal random variable. An ad hoc alternative would be use a *t*-distribution with $n-p$ degrees of freedom as the reference distribution for small or moderate size samples. If e_i is not normal, the Gauss inequality (Pukelsheim 1994, Weisstein 2006) is useful for setting a cutoff value.

Gauss Inequality: If a distribution has a single mode at μ_0, then

$P\left\{\left|x-\mu_0\right| > \lambda\tau\right\} \leq \frac{4}{9\lambda^2}$, where $\tau^2 \equiv \sigma^2 + \left(\mu-\mu_0\right)^2$.

According to the model 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 σ, we can use either 2 as a loose cutoff or 3 as a strict one to identify outlying residuals, depending on analysts' preference. Note that it is not feasible to standardize using the robust estimate of $V_M\left(e_i\right)$ discussed in Section 3.5.2. The robust estimate of $se\left(e_i\right)$

would be $\sqrt{e_i^2}$, which would create a degenerate case for the standardized residual.

For multistage sampling and its corresponding model (3.9), the residual can still be justified after rescaled by its appropriately estimated standard error. The residual vector for sample cluster *i* is $\mathbf{e}_i = \mathbf{Y}_i - \sum_{i' \in s} \mathbf{H}_{ii'} \mathbf{Y}_{i'}$ and its variance-covariance matrix is (3.12). Within a cluster *i*, assume the residuals $\mathbf{e}_i$ are jointly normally distributed. Then its *k*th element e_{ik} is marginally normally distributed with mean zero and variance $\left[V_M\left(\mathbf{e}_i\right)\right]_{kk} \cong \sigma^2$ if model (3.9) is correct. After obtaining the estimates of $\left(1-\rho\right)\sigma^2$ and $\rho\sigma^2$ described in Section 3.5.3, we can divide e_{ik} by $\hat{\sigma} = \sqrt{\hat{P} + \frac{\hat{Q} - \hat{P}}{\hat{D}}}$ to standardize it, where $\hat{P}$, $\hat{Q}$, and $\hat{D}$ are also defined in Section 3.5.3. As for single-stage sampling, use of the robust estimate of $V_M(\mathbf{Y}_i)$ is not feasible for standardization because it involves only $\mathbf{e}_i \mathbf{e}_i^T$.

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

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 $\hat{\beta}_j$ in a multiple regression model. This plot is useful for graphically detecting influential points and outliers, so that we can use it as a good alternative and supplement to DFBETAS, etc. Korn and Graubard (1999) illustrated the use of these

plots with survey data. Let $\mathbf{X}(-j)$ be $n\times(p-1)$ matrix formed from the data matrix, $\mathbf{X}$, by removing its *j*th column $\mathbf{x}^j$. Further let $\mathbf{u}_j$ and $\mathbf{v}_j$ be the residuals that result from regressing $\mathbf{Y}$ and $\mathbf{x}^j$ on $\mathbf{X}(-j)$ using survey weights. It is known that $\hat{\beta}_j$, the *j*th regression coefficient of a multiple regression model, is the same as the slope coefficient of the weighted two-variate regression of $\mathbf{u}_j$ on $\mathbf{v}_j$. The added variable plot is defined as a scatter plot of $\mathbf{u}_j$ against $\mathbf{v}_j$ along with their simple linear regression line. For survey data it can be drawn as a bubble plot with each bubble representing an observation and its area proportional to the sample weight. By itself the plot is not able to precisely measure how severely an observation is different from others, but 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.

3.6.2 DFBETAS

Taking the sampling weights $\mathbf{W}$ into consideration,

$$DFBETA_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = \frac{\mathbf{A}^{-1}\mathbf{x}_i e_i w_i}{1-h_{ii}} \quad \text{(see, e.g., Valliant, et al. 2000)}$$

with $\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ for single-stage sampling, or $DFBETA_{ik} = \frac{\mathbf{A}^{-1}\mathbf{x}_{ik} e_{ik} w_{ik}}{1-h_{ik,ik}}$ for clustered sampling, where $h_{ik,ik} = \mathbf{x}_{ik}^T\mathbf{A}^{-1}\mathbf{x}_{ik}w_{ik}$, with subscript ik indicating the *k*th unit within the *i*th cluster, is the *k*th diagonal element on the matrix $\mathbf{H}_{ii} = \mathbf{X}_i\mathbf{A}^{-1}\mathbf{X}_i^T\mathbf{W}_i$ (defined in Section 3.5.3). Although the formulas for the DFBETA statistic look very much like the one in the OLS case, they have differences in both numerator and denominator because sample weights are involved in the leverages and residuals. However, the formulas have exactly the same form as the one for WLS with weights inversely proportional to model variances. To create a complex sample version of DFBETAS, we need to divide DFBETA by an estimate of the standard error of $\hat{\boldsymbol{\beta}}$ that

accounts for unequal weighting, stratification, clustering, and other design complexities.

(1) Single Stage Sampling

By knowing $DFBETA_{ij} = \dfrac{\left(\mathbf{A}^{-1}\mathbf{x}_i e_i w_i\right)_j}{1-h_{ii}} = \dfrac{c_{ji}e_i}{1-h_{ii}}$ and the variance estimator of $\hat{\beta}_j$, $v_M\left(\hat{\beta}_j\right) = \hat{\sigma}^2\left[\mathbf{A}^{-1}\left(\sum_{i'=1}^{n} w_{i'}^2 \mathbf{x}_{i'}\mathbf{x}_{i'}^T\right)\mathbf{A}^{-1}\right]_{jj} = \hat{\sigma}^2\sum_{i'=1}^{n} c_{ji'}^2$ where $c_{ji'}$ is an element of matrix $\mathbf{C}$ defined in Section 3.5.1, under model (2.1), we are able to construct a scaled statistic DFBETAS as in the OLS case. We propose a specification of DFBETAS statistic as follows:

$$DFBETAS_{ij} = \frac{c_{ji}e_i/\left(1-h_{ii}\right)}{\sqrt{v_M\left(\hat{\beta}_j\right)}} = \frac{c_{ji}}{\sqrt{\sum_{i'=1}^{n} c_{ji'}^2}}\cdot\frac{e_i}{\hat{\sigma}}\cdot\frac{1}{1-h_{ii}}.$$

Using the order conditions $c_{jk} = O\left(n^{-1}\right)$ and $h_{ii} = O\left(n^{-1}\right)$, we rewrite the DFBETAS statistic as the approximate product of two terms, $DFBETAS_{ij} \doteq O\left(n^{-1/2}\right)\cdot N(0,1)$. The first term, with an order of $n^{-1/2}$, can be approximated by $n^{-1/2}$ when the sampled units have similar $\mathbf{X}$ values and weights. An observation i may be identified as influential on the estimation of $\hat{\beta}_j$ if $\left|DFBETAS_{ij}\right| \geq \dfrac{2}{\sqrt{n}}$. Moreover, the model robust sandwich estimator $v_W\left(\hat{\beta}_j\right)$ and the linearization estimator $v_L\left(\hat{\beta}_j\right)$ can be used to replace $v_M\left(\hat{\beta}_j\right)$ to guard against the possibility that the underlying model deviates from the working model. An ad hoc alternative would be to use a cutoff of $t_{0.025}(n-p)/\sqrt{n}$ where $t_{0.025}(n-p)$ is the 97.5 percentile of the t-distribution with

$n-p$ degree of freedom. We can also use $\left|DFBETAS_{ij}\right| \geq \frac{3}{\sqrt{n}}$ as a more generous criterion if the normality of the residuals does not hold.

(2) Multiple Stage Sampling

In the case of a multi-stage complex sampling design, the DFBETAS statistic is constructed in a similar way as the one in the case of single-stage sampling, except that the variance estimator of $\hat{\beta}_j$ needs to be replaced by $v_M\left(\hat{\beta}_j\right)$ from model (3.9). Since

$$\begin{aligned} v_M\left(\hat{\boldsymbol{\beta}}\right) &= \sum_s \mathbf{A}^{-1}\mathbf{X}_i\mathbf{W}_i\left[\widehat{(1-\rho)\sigma^2}\mathbf{I}_{m_i} + \widehat{\rho\sigma^2}\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\right]\mathbf{W}_i\mathbf{X}_i^T\mathbf{A}^{-1} \\ &= \hat{\sigma}^2\sum_s \mathbf{C}_i\left[(1-\hat{\rho})\mathbf{I}_{m_i} + \hat{\rho}\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\right]\mathbf{C}_i^T \end{aligned}$$

where $\mathbf{C}_i$ is a $p \times m_i$ submatrix of $\mathbf{C}$ and defined as $\mathbf{C}_i = \mathbf{A}^{-1}\mathbf{X}_i\mathbf{W}_i$ with (jk)th element $c_{j,ik}$ ($j = 1,\ldots,p; k = 1,\ldots,m_i$), we have

$$\begin{aligned} v_M\left(\hat{\boldsymbol{\beta}}_j\right) &= \hat{\sigma}^2\sum_s \left(c_{j,i1}\ldots c_{j,im_i}\right)\begin{pmatrix} 1 & & \hat{\rho} \\ & \ddots & \\ \hat{\rho} & & 1 \end{pmatrix}\left(c_{j,i1}\ldots c_{j,im_i}\right)^T \\ &= \hat{\sigma}^2\sum_s\left(\sum_{k=1}^{m_i} c_{j,ik}^2 + \hat{\rho}\sum_{k\neq k'}^{m_i} c_{j,ik}c_{j,ik'}\right). \end{aligned}$$

The constructed DFBETAS statistic is specified as

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

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 $\frac{1}{\sqrt{n\bar{m}\left[1+\hat{\rho}\left(\bar{m}-1\right)\right]}}$ with $\bar{m}=\frac{1}{n}\sum_{s} m_i$. Note that $1+\hat{\rho}\left(\bar{m}-1\right)$ is the estimated design effect (Kish, 1995). We propose that the cutoff value for DFBETAS statistics can be set as $\frac{2}{\sqrt{n\bar{m}\left[1+\hat{\rho}\left(\bar{m}-1\right)\right]}}$ or $\frac{3}{\sqrt{n\bar{m}\left[1+\hat{\rho}\left(\bar{m}-1\right)\right]}}$. There are two options to obtain the $\hat{\rho}$ in the above cutoffs: 1) Estimate ρ using (3.10); 2) Estimate ρ from $1+\hat{\rho}\left(\bar{m}-1\right)=\frac{v_M\left(\hat{\beta}_j\right)}{v_{SRS}\left(\hat{\beta}_j\right)}$. If an individual observation is greatly distinguished from the other observations in the sample, it might amplify the DFBETAS statistics and make it exceed the cutoff in two possible ways: 1) through an outlying residual; 2) through an outlying leverage. Since the single-stage sampling can be viewed as a special case of the multistage complex sampling in which there is only unit within each sampled PSU, or $\bar{m}=1$, the above cutoff boils down to $\frac{2}{\sqrt{n}}$ or $\frac{3}{\sqrt{n}}$ with n defined as the sample size, which corresponds to what we have obtained in case (1). Note that the model based variance estimator v_M can be replaced by the sandwich estimator v_W and the linearization estimator v_L to protect against the deviation from model (3.9) and to facilitate design based interpretations. This replacement can also be applied to the diagnostic statistics that will be discussed below.

3.6.3 DFFITS

Multiplying the DFBETA statistic by the $\mathbf{x}_i^T$ vector, we obtain the measure of change in the *i*th fitted values due to the deletion of the *i*th observation, $DFFIT_i=\hat{Y}_i-\hat{Y}_i(i)=\mathbf{x}_i^T\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)\right)=\frac{h_{ii}e_i}{1-h_{ii}}$ for the single-stage sampling. The model

variance of $\hat{Y}_i$ is $V_M\left(\hat{Y}_i\right)=\sigma^2\left(\mathbf{HH}^T\right)_{ii}=\sigma^2\sum_{i'}h_{ii'}^2$, which is estimated by $v_M\left(\hat{Y}_i\right)=\hat{\sigma}^2\sum_{i'}h_{ii'}^2$. In OLS, $\sum_{i'}h_{ii'}^2=h_{ii}$ 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 (2.1), DFFIT$_i$ is divided by the square root of $v_M\left(\hat{Y}_i\right)$ and rearranged as follows:

$$\begin{aligned}DFFITS_i&=\frac{h_{ii}e_i/\left(1-h_{ii}\right)}{\sqrt{v_M\left(\hat{Y}_i\right)}}\\&=\frac{h_{ii}e_i/\left(1-h_{ii}\right)}{\sqrt{\hat{\sigma}^2\sum_{i'}h_{ii'}^2}}\\&=\frac{h_{ii}}{1-h_{ii}}\frac{e_i}{\hat{\sigma}}\frac{1}{\sqrt{\sum_{i'}h_{ii'}^2}}.\end{aligned}$$

When the sample weights do not have a large variation, we have $\sum_{i'}h_{ii'}^2\approx h_{ii}$. Because the mean of the leverages is $\frac{p}{n}$, we can set the cutoff value to be $2\sqrt{\frac{p}{n}}$ for using DFFITS to determine the influential observations.

If a sample is drawn from a complex clustering design, or, the working model considers clustering, the DFFIT statistic becomes $DFFIT_{ik}=\frac{h_{ik,ik}e_{ik}}{1-h_{ik,ik}}$. The variance of the predicted value is estimated as

$$\begin{aligned}v_M\left(\hat{Y}_{ik}\right)&=\mathbf{x}_{ik}^Tv_M\left(\hat{\boldsymbol{\beta}}\right)\mathbf{x}_{ik}=\mathbf{x}_{ik}^T\sum_s\mathbf{A}^{-1}\mathbf{X}_i^T\mathbf{W}_i\left(\widehat{\left(1-\rho\right)\sigma^2}\mathbf{I}_{m_i}+\widehat{\rho\sigma^2}\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\right)\mathbf{W}_i\mathbf{X}_i\mathbf{A}^{-1}\mathbf{x}_{ik}\\&=\hat{\sigma}^2\sum_{i'\in s}\left(h_{ik,i'1}\cdots h_{ik,i'm_{i'}}\right)\left(\widehat{\left(1-\rho\right)\sigma^2}\mathbf{I}_{m_i}+\widehat{\rho\sigma^2}\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\right)\left(h_{ik,i'1}\cdots h_{ik,i'm_{i'}}\right)^T\\&=\hat{\sigma}^2\sum_{i'\in s}\left(\sum_{k'=1}^{m_{i'}}h_{ik,i'k'}^2+\hat{\rho}\sum_{k''\neq k'}^{m_{i'}}h_{ik,i'k'}h_{ik,i'k''}\right).\end{aligned}$$

.

Therefore, the DFFITS statistic is formulated as

$$DFFITS_{ik} = \frac{h_{ik,ik}e_{ik}\big/\left(1-h_{ik,ik}\right)}{\sqrt{v_M\left(\hat{Y}_{ik}\right)}}$$

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

where $h_{ik,i'k'} = \mathbf{x}_{ik}^T\mathbf{A}^{-1}\mathbf{x}_{i'k'}w_{i'k'}$ is an element of $\mathbf{H}_{ii'} = \mathbf{X}_i\mathbf{A}^{-1}\mathbf{X}_{i'}^T\mathbf{W}_{i'}$. 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,

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

The cutoff for the DFFITS statistic is determined to be $2\sqrt{\frac{p}{n\bar{m}\left[1+\hat{\rho}\left(\bar{m}-1\right)\right]}}$ when ρ is estimated appropriately. Naturally, $v_M\left(\hat{\boldsymbol{\beta}}\right)$ in the formula can be replaced by $v_W\left(\hat{\boldsymbol{\beta}}\right)$ from model (3.11) or $v_L\left(\hat{\boldsymbol{\beta}}\right)$ to accommodate a general situation. We can also consider $3\sqrt{\frac{p}{n\bar{m}\left[1+\hat{\rho}\left(\bar{m}-1\right)\right]}}$ as a less strict cutoff.

3.6.4 Distance Measure (Extended and Modified Cook's Distance)

Under model (3.5) $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim ind\left(0,\psi_i\right)$, according to Theorem 3.17, Theorem 3.12, and Corollary 1.3 in Shao (1999), under some regularity conditions, we have $\boldsymbol{\Sigma}^{-1/2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d} N\left(\mathbf{0},\mathbf{I}\right)$, where $\boldsymbol{\Sigma} = V_M\left(\hat{\boldsymbol{\beta}}\right)$, as in expression (3.6). Since $v_W\left(\hat{\boldsymbol{\beta}}\right)$ is a consistent estimator of $V_M\left(\hat{\boldsymbol{\beta}}\right)$, the statistics constructed by replacing $\boldsymbol{\Sigma}$ by $v_W\left(\hat{\boldsymbol{\beta}}\right)$ have the same limiting distributions:

$$\left[v_W\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1/2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d} N_p(\mathbf{0},\mathbf{I})$$

$$\text{and } \left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T\left[v_W\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}\chi^2(p). \tag{3.19}$$

Under the linear model considering stratification and cluster samples, we can draw similar conclusions because $\hat{\boldsymbol{\beta}}=\sum_h\sum_{i\in s_h}\sum_{k\in s_{hi}}\mathbf{A}^{-1}\mathbf{x}_{hik}w_{hik}Y_{hik}=\sum_{h,i\in s_h}\mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{Y}_{hi}$, which is the sum of $\sum_h n_h$ weighted cluster totals and the clusters are assumed to be independently selected. In this case the sandwich variance estimator $v_W\left(\hat{\boldsymbol{\beta}}\right)$ is formulated as in (3.7) to take account of the correlations within the clusters. Alternatively, $v_L\left(\hat{\boldsymbol{\beta}}\right)$ in (3.15), which is asymptotically equivalent to $v_W\left(\hat{\boldsymbol{\beta}}\right)$, can be used. Equations (3.19) still hold if the true model parameter $\boldsymbol{\beta}$ is replaced by the finite population parameter $\mathbf{B}$ under some regularity conditions and some sampling designs (Fuller 1975, 2002).

The classical Wald statistic, based on the second expression in (3.19) which approaches a chi-square distribution, is often used to test a set of hypotheses about slope coefficients for multiple linear regression analyses. The use of this statistic was also introduced for tests on regression coefficients of complex survey data. For example, the Wald chi-square statistic for testing the hypothesis $H_0:\mathbf{B}=\mathbf{B}_0$ is

$$WD=\left(\hat{\boldsymbol{\beta}}-\mathbf{B}_0\right)^T\left[v\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\hat{\boldsymbol{\beta}}-\mathbf{B}_0\right),$$

where $\mathbf{B}_0$ is the hypothesized value of the finite population parameter vector $\mathbf{B}$ and $v\left(\hat{\boldsymbol{\beta}}\right)$ is a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$, computed by approaches such as balanced half-sample replication, Taylor series linearization estimator $v_L\left(\hat{\boldsymbol{\beta}}\right)$, or sandwich estimator $v_W\left(\hat{\boldsymbol{\beta}}\right)$. Under $H_0:\mathbf{B}=\mathbf{B}_0$, WD is asymptotically distributed as a chi-square random variable with p degrees of freedom. The Wald F statistic is obtained by dividing WD by p: $F_W=WD/p$. Under H_0, an ad hoc approach is to treat F_W as an F random variable with p and r degrees of freedom, where r is the degrees of freedom associated with $v\left(\hat{\boldsymbol{\beta}}\right)$. For multistage designs r is usually taken

to be the number of PSU's minus the number of first stage strata. In order for the Wald statistic WD to perform properly, $v\left(\hat{\boldsymbol{\beta}}\right)$ must be a consistent estimator of the true variance, $V\left(\hat{\boldsymbol{\beta}}\right)$. This variance can be computed with respect to either a design or a model. If an inconsistent variance estimator, e.g., the OLS variance estimator, is used, then WD will not be χ_p^2 distributed even in large samples.

Korn and Graubard (1990) demonstrate that an adjusted Wald F statistic, $F_{ADJWF} = \frac{n-p+1}{np} WD$, can be a real improvement over the asymptotically correct chi-square distribution when the number of regression coefficients approaches the degrees of freedom available from the variance estimation. The F statistic is distributed with p and $n-p+1$ degrees of freedom under H_0, where n is the number of sample clusters. Other alternatives are also available, such as Rao-Scott first-order and second-order corrections (Rao & Scott, 1980) and Fay's replication approach (Fay, 1985).

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

$$ED_i = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)\right)^T \left[v_W\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)\right). \tag{3.20}$$

Since the method of calibrating the Cook's Distance is obtained by analogy to a confidence ellipsoid, the newly created statistic ED_i can be mapped to a Chi-square distribution. If ED_i were exactly equal to the $(1-\alpha)\times 100\%$ level of the Chi-square distribution with p degrees of freedom, then the deletion of the ith case would move the estimate of $\boldsymbol{\beta}$ to the edge of a $(1-\alpha)\times 100\%$ confidence ellipsoid based on the complete data. A large value of this quadratic term indicates that the ith observation is likely to be influential in determining the joint inferences about all the parameters in the regression model. The variance estimator $v_W\left(\hat{\boldsymbol{\beta}}\right)$ can be replaced by the linearization variance estimator $v_L\left(\hat{\boldsymbol{\beta}}\right)$ since both of them are design and model consistent.

Another formulation of the extended Cook's Distance can be derived from the Wald F statistic as $ED_i' = \frac{n-p+1}{np}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)\right)^T\left[v_W\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)\right)$ and its value can be compared with an F distribution.

Like the Cook's Distance, the proposed extended Cook's Distance statistic is related to the sample size in order of magnitude. However, the F and Chi-square statistics do not change very much when the sample size exceeds 100 or more. Therefore, very few observations can be identified to be influential in that case even if the small percentiles of F and Chi-square statistics are adopted as cutoffs. Following Atkinson (1982), we modify the proposed extended Cook's Distance to solve this problem.

Suppose the sample is drawn from a single-stage design and the working model is (2.1). Then

$$\begin{aligned} ED_i &= \left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)\right)^T\left[v_M\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)\right) \\ &= \left(\frac{e_i}{\hat{\sigma}}\right)^2 \frac{1}{\left(1-h_{ii}\right)^2} w_i \mathbf{x}_i^T \mathbf{A}^{-1}\left[\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{W}\mathbf{X}\mathbf{A}^{-1}\right]^{-1}\mathbf{A}^{-1}\mathbf{x}_i w_i \qquad (3.21) \\ &= \left(\frac{e_i}{\hat{\sigma}}\right)^2 \frac{1}{\left(1-h_{ii}\right)^2} w_i \mathbf{x}_i^T \left[\mathbf{X}^T\mathbf{W}\mathbf{W}\mathbf{X}\right]^{-1}\mathbf{x}_i w_i. \end{aligned}$$

Based on the assumptions in Section 3.5.1, we know that

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

and this quantity has a mean of p/n. Hence, we suggest that an analyst take the square root of the extended Cook's D statistic and rescale the root by $\left(n/p\right)^{-1/2}$. The modified statistic $MD_i = \sqrt{nED_i/p}$, called the modified Cook's D, can be judged in terms of a standard normal distribution, or in other words, we can use 2 as the cutoff value. If the assumption of normality is violated, we can use a more generous cutoff, 3, in terms of Gauss Inequality.

For a cluster sample it is convenient to use $v_M\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2\mathbf{X}^T\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\mathbf{X}$ rather than the equivalent form given in Section 3.6.2 and 3.6.3. The matrix $\boldsymbol{\Phi}$ is block diagonal with 1 on the diagonal and ρ off the diagonal in each block (cluster). The dimension of

block i is $m_i \times m_i$. If we assume the working model is (3.9), the modified Cook's D statistic becomes

$$\begin{aligned} ED_{ik} &= \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(ik)\right)^T \left[v_M\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(ik)\right) \\ &= \left(\frac{e_{ik}}{\hat{\sigma}}\right)^2 \frac{1}{\left(1 - h_{ik,ik}\right)^2} w_{ik} \mathbf{x}_{ik}^T \left[\mathbf{X}^T \mathbf{W} \boldsymbol{\Phi} \mathbf{W} \mathbf{X}\right]^{-1} \mathbf{x}_{ik} w_{ik} \end{aligned} \tag{3.22}$$

where $\hat{\boldsymbol{\beta}}(ik)$ is the parameter estimate after deleting unit *k* in cluster *i*. If the number of units within each sampled PSU, m_i, is bounded,

$$w_{ik} \mathbf{x}_{ik}^T \left[\mathbf{X}^T \mathbf{W} \boldsymbol{\Phi} \mathbf{W} \mathbf{X}\right]^{-1} \mathbf{x}_{ik} w_{ik} = O\left(n^{-1}\right),$$

where n is the number of sampled PSUs (see proof below), and the value of this expression is approximately equal to $p\left[n\bar{m}\left(1 + \hat{\rho}\left(\bar{m} - 1\right)\right)\right]^{-1}$ 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 $\frac{2\sqrt{p}}{\sqrt{n\bar{m}\left[1 + \hat{\rho}\left(\bar{m} - 1\right)\right]}}$ or $\frac{3\sqrt{p}}{\sqrt{n\bar{m}\left[1 + \hat{\rho}\left(\bar{m} - 1\right)\right]}}$. Also, we can define

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

and compare it to 2 or 3.

The following is the proof of $w_{ik} \mathbf{x}_{ik}^T \left[\mathbf{X}^T \mathbf{W} \boldsymbol{\Phi} \mathbf{W} \mathbf{X}\right]^{-1} \mathbf{x}_{ik} w_{ik} = O\left(n^{-1}\right)$:

We have $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}$, $\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 & 0 \\ \ddots & \\ 0 & \mathbf{W}_n \end{pmatrix} = O\left(\frac{N}{n}\right)$ where $\mathbf{W}_i$ is a diagonal matrix of weights in cluster *i*, and $\boldsymbol{\Phi} = blkdiag \begin{pmatrix} 1 & \rho \\ \ddots & \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Phi}_1 & 0 \\ \ddots & \\ 0 & \boldsymbol{\Phi}_n \end{pmatrix} = O(1)$ where

$\boldsymbol{\Phi}_i = \begin{pmatrix} 1 & \rho \\ \ddots & \\ \rho & 1 \end{pmatrix}_{m_i \times m_i}$, $i = 1, \ldots, n$.

Therefore,

$$\mathbf{X}^T\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\mathbf{X} = \left(\mathbf{X}_1^T \cdots \mathbf{X}_n^T\right)\begin{pmatrix}\mathbf{W}_1 & 0\\ \ddots & \\ 0 & \mathbf{W}_n\end{pmatrix}\begin{pmatrix}\boldsymbol{\Phi}_1 & 0\\ \ddots & \\ 0 & \boldsymbol{\Phi}_n\end{pmatrix}\begin{pmatrix}\mathbf{W}_1 & 0\\ \ddots & \\ 0 & \mathbf{W}_n\end{pmatrix}\begin{pmatrix}\mathbf{X}_1\\ \vdots \\ \mathbf{X}_n\end{pmatrix},$$

$$= \sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{W}_i\boldsymbol{\Phi}_i\mathbf{W}_i\mathbf{X}_i = O\left(n\frac{N^2}{n^2}\right) = O\left(\frac{N^2}{n}\right)$$

$$w_{ik}\mathbf{x}_{ik}^T\left[\mathbf{X}^T\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\mathbf{X}\right]^{-1}\mathbf{x}_{ik}w_{ik} = O\left(\frac{N}{n}\right)O\left(\frac{n}{N^2}\right)O\left(\frac{N}{n}\right) = O\left(n^{-1}\right).$$

3.6.5 Discussion

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. Sometimes, if needed, we can also use the estimate of $Var\left(\hat{\boldsymbol{\beta}}(i)\right)$ because it is sensitive to the deletion of observation i. The same cutoffs can be applied since both $Var\left(\hat{\boldsymbol{\beta}}(i)\right)$ and $Var\left(\hat{\boldsymbol{\beta}}\right)$ are of the same order, n^{-1}.

The proof of the above statements is as follows:

We have

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(i) + \frac{\mathbf{A}^{-1}\mathbf{x}_i^T e_i w_i}{1-h_{ii}} = \hat{\boldsymbol{\beta}}(i) + \frac{\mathbf{A}^{-1}\mathbf{x}_i^T\left(Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)w_i}{1-h_{ii}}.$$

Move the term including $\hat{\boldsymbol{\beta}}$ to the left hand side of the equation to obtain

$$\left(\mathbf{I} + \frac{\mathbf{A}^{-1}\mathbf{x}_i^T\mathbf{x}_i w_i}{1-h_{ii}}\right)\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(i) + \frac{\mathbf{A}^{-1}\mathbf{x}_i^T Y_i w_i}{1-h_{ii}}.$$

Using the order of magnitude analysis, the above analysis can be simplified as

$$\left[\mathbf{I} + O\left(n^{-1}\right)\right]\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(i) + O\left(n^{-1}\right)Y_i.$$

Take the variances for both sides:

$$O(1)\cdot Var\left(\hat{\boldsymbol{\beta}}\right)=Var\left(\hat{\boldsymbol{\beta}}(i)\right)+O\left(n^{-2}\right)\cdot\psi_i=Var\left(\hat{\boldsymbol{\beta}}(i)\right)+O\left(n^{-2}\right).$$

Since $Var\left(\hat{\boldsymbol{\beta}}\right)=\sum_{k=1}^{n}c_{jk}^2\psi_i=O\left(n^{-1}\right)$, we conclude $Var\left(\hat{\boldsymbol{\beta}}(i)\right)$ is of order n^{-1} and their estimates should have the same orders, too.

The determination of influential observations usually involves choosing reasonable cutoffs which are suitable for the problem at hand and guided by statistical theory. However, some diagnostic statistics such as leverages are not directly related to natural standard error scaling. Moreover, under some occasions, deriving a design-based distribution for corresponding diagnostics does not appear to be possible. There are some criteria seem useful for these cases. What Belsley, Kuh, and Welsch (1980) refer to as internal scaling means to "define extreme values of a diagnostic measure relative to the *weight of the evidence* provided by the given diagnostic series itself." Suppose we generate a series of size *n* by calculating some diagnostic statistic, say leverages. The interquartile range, defined as $IQR\equiv Q3-Q1$, can be computed for that series and extreme leverages are indicated as those exceeding (7/2)IQR. It is convenient to use interquartile range for influence identification in the absence of a more exact distribution theory since it provides a more robust estimate of spread, especially when the underlying distribution is non-Gaussian or highly skewed.

Another useful and intuitive way of catching outliers is to pay attention to the gap which appears in the series of a diagnostic measure. Usually, it is worthy of notice if the large majority of the elements in a diagnostic series have similar values, but small fractions of observations are noticeably larger or smaller than the others. However, there is lack of theoretical support to determine the largeness of a gap.

In summary, the influence analysis is based on theoretically justified diagnostic measures and their cutoffs, but sometimes the criteria can be flexible and case-specific.

Chapter 4: Identification of Influential Groups of Observations

In Chapter 3, we have presented various diagnostic techniques for identifying influential observations that have been based on the deletion of a single unit. However, such techniques will not always be successful. Sometimes they may not be able to identify any influential cases since a single observation is less likely to have a significant effect on parameter estimation when the data set is large. Even if some influential points are located, one outlier can mask the effect of another. It is necessary, therefore, to develop techniques that examine the potentially influential effects of subsets or groups of observations. This is especially important in a large survey data set where a few individual units may have a limited effect but a group of units may be more important. For example, in a clustered survey using geographic primary sampling units (PSUs) and personal interviews, common practice is to use one or two data collectors per PSU. If an interviewer produces correlated data among units with a level different from the average, residuals for the units done by that interviewer may be consistently positive or negative and in some cases extreme.

4.1 Multiple-Case Deletion

In the conventional diagnostics, Belsley, Kuh, and Welsch (1980) presented examples of a natural multiple-row generalization of DFBETA and DFFIT. For example, a

measure of the change in coefficients, $\frac{\left|\mathbf{b}_j - \mathbf{b}_j\left(\mathbf{D}_m\right)\right|}{\text{scale}}$, where $\mathbf{D}_m$ is a deletion set of size m, and "scale" indicates some appropriate measure of standard error. If the fitted values are of interest, the appropriate measure becomes $\frac{\left|\mathbf{x}_i\left[\mathbf{b} - \mathbf{b}\left(\mathbf{D}_m\right)\right]\right|}{\text{scale}}$, where $\mathbf{x}_i$ is the vector of covariates for unit i. To avoid multiple computational tasks for each deletion set, the quadratic form $MDFFIT \equiv \left[\mathbf{b} - \mathbf{b}\left(\mathbf{D}_m\right)\right]^T \mathbf{X}^{\mathbf{T}}\mathbf{X}\left[\mathbf{b} - \mathbf{b}\left(\mathbf{D}_m\right)\right]$ can be considered as a summary measure. Meanwhile, they pointed out that as *m* becomes large, the heavy computational burden and the difficulty of finding the starting subset will tend to limit the applications of those techniques. They suggested that a stepwise approach can provide useful information at relatively low cost.

The stepwise approach starts by forming the initial subset $\mathbf{D}_m$ of size *m*, say $m = 2$, using the observations with the two largest $\left|\text{DFFIT}\right|$ or $\left|\text{DFFITS}\right|$ computed using the delete-one method. If the two largest values of $\left|\mathbf{x}_i\left[\mathbf{b} - \mathbf{b}\left(\mathbf{D}_m\right)\right]\right|$, where $i \in s$, the full sample, do not have their indexes *i* contained in $\mathbf{D}_2$, $\mathbf{D}_2$ is reconstructed consisting of the indexes for the two largest. This procedure proceeds until the indexes of the two observations in $\mathbf{D}_2$ coincide with the two largest values of $\left|\mathbf{x}_k\left[\mathbf{b} - \mathbf{b}\left(\mathbf{D}_m\right)\right]\right|$. Then a starting set $\mathbf{D}_3$ is found using the three largest values of $\left|\mathbf{x}_k\left[\mathbf{b} - \mathbf{b}\left(\mathbf{D}_m\right)\right]\right|$ from the previous iteration for $m = 2$. And the overall process continues until the subset size reaches the desired *m*. For large datasets, Belsley, Kuh, and Welsch (1980) recommend a single-row deletion analysis coupled with the partial-regression leverage plots and stepwise multiple-row methods to enhance efficiency and effectiveness.

Some of the single-row deletion diagnostics for survey data generalized to their multiple-row deletion versions are summarized as follows:

(1) $DFBETA_D \equiv \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)} = \mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D$;

(2) $DFFIT_D \equiv \mathbf{X}_D\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)}\right) = \mathbf{X}_D\mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D = \mathbf{H}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D$;

(3) $MDFFIT_D \equiv \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)}\right)^T \mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{X}_{(D)}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)}\right)$

$$= \mathbf{e}_D^T\mathbf{W}_D\mathbf{X}_D\left(\mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{X}_{(D)}\right)^{-1}\mathbf{X}_D^T\mathbf{W}_D\mathbf{e}_D;$$

(4) $ED_D \equiv \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)}\right)^T\left[v\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)}\right)$, called the extended Cook's Distance here as it was in Chapter 3;

where D is a set of indices which denote the observations that will be deleted from the regression, so that we have

$$\mathbf{X}_D = \left(\mathbf{x}_i\right)^T,\ i \in D;$$

$$\mathbf{W}_D = diag\left(w_i\right),\ i \in D;$$

$$\mathbf{e}_D = \left(e_i\right)^T,\ i \in D;$$

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

where $\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ as in Chapter 3, and

$$\mathbf{X}_{(D)} = \left(\mathbf{x}_i\right)^T,\ i \notin D;$$

$$\mathbf{W}_{(D)} = diag\left(w_i\right),\ i \notin D;$$

$$\hat{\boldsymbol{\beta}}_{(D)} = \left(\mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{X}_{(D)}\right)^{-1}\mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{Y}_{(D)}.$$

The derivations of (1) and (3) are given below. Shao (1988) also covers the idea of

deleting a group of units when using the jackknife. His article does address regression, but the development below is new and covers the problems that are specifically of interest here.

Since it is not easy to determine the distribution-based cutoff values for above generalized statistics, they can be evaluated using the scaled measures relative to their maxima, instead. The extended Cook's Distance can be evaluated using the variance estimate for $\hat{\boldsymbol{\beta}}_{(D)}$ instead of $\hat{\boldsymbol{\beta}}$ because it may be sensitive to the exclusion of the influential group when the size of the deletion group is relatively large.

Proofs of (1) and (3):

(1) $DFBETA_D = \mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D$

To verify this formula, we use the result from Schott (1997, Theorem 1.7) that for conformable matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$,

$$\left(\mathbf{A}+\mathbf{CBD}\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1}+\mathbf{DA}^{-1}\mathbf{C}\right)^{-1}\mathbf{DA}^{-1},$$

assuming $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ exist. With $\mathbf{A}=\mathbf{X}^T\mathbf{WX}$, $\mathbf{B}=\mathbf{I}$, $\mathbf{C}=-\mathbf{X}_D^T\mathbf{W}_D$, and $\mathbf{D}=\mathbf{X}_D$, we have

$$\left(\mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{X}_{(D)}\right)^{-1} = \left(\mathbf{X}^T\mathbf{WX}-\mathbf{X}_D^T\mathbf{W}_D\mathbf{X}_D\right)^{-1} = \mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{X}_D\mathbf{A}^{-1} \tag{4.1}$$

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(D)} &= \left(\mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{X}_{(D)}\right)^{-1}\mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{Y}_{(D)} \\
&= \left[\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{X}_D\mathbf{A}^{-1}\right]\left(\mathbf{X}^T\mathbf{W}\mathbf{Y} - \mathbf{X}_D^T\mathbf{W}_D\mathbf{Y}_D\right) \\
&= \hat{\boldsymbol{\beta}} + \mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{X}_D\hat{\boldsymbol{\beta}} - \mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\mathbf{Y}_D - \mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{H}_D^T\mathbf{Y}_D \\
&= \hat{\boldsymbol{\beta}} + \mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\left(\hat{\mathbf{Y}}_D - \left(\mathbf{I}-\mathbf{H}_D\right)\mathbf{Y}_D - \mathbf{H}_D\mathbf{Y}_D\right) \\
&= \hat{\boldsymbol{\beta}} - \mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D.
\end{aligned}$$

(3) $$MDFFIT_D = \mathbf{e}_D^T\mathbf{W}_D\mathbf{X}_D\left(\mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{X}_{(D)}\right)^{-1}\mathbf{X}_D^T\mathbf{W}_D\mathbf{e}_D.$$

Using the expression for $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)}$ implied by $DFBETA_D$, $MDFFIT_D$ can be rewritten as

$$\begin{aligned}
MDFFIT_D &\equiv \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)}\right)^T\mathbf{X}_{(D)}^T\mathbf{W}_{(D)}\mathbf{X}_{(D)}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(D)}\right) \\
&= \mathbf{e}_D^T\left[\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\right]^T\mathbf{W}_D\mathbf{X}_D\mathbf{A}^{-1}\left(\mathbf{A} - \mathbf{X}_D^T\mathbf{W}_D\mathbf{X}_D\right)\mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D \\
&= \mathbf{e}_D^T\left[\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\right]^T\mathbf{W}_D\mathbf{X}_D\mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D \\
&\quad - \mathbf{e}_D^T\left[\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\right]^T\mathbf{W}_D\mathbf{X}_D\mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\mathbf{X}_D\mathbf{A}^{-1}\mathbf{X}_D^T\mathbf{W}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D \\
&= \mathbf{e}_D^T\left[\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\right]^T\mathbf{W}_D\mathbf{H}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D \\
&\quad - \mathbf{e}_D^T\left[\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\right]^T\mathbf{W}_D\mathbf{H}_D\mathbf{H}_D\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{e}_D \\
&= \mathbf{e}_D^T\left[\left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\right]^T\mathbf{W}_D\mathbf{H}_D\mathbf{e}_D.
\end{aligned}$$

Next, transposing the expression above and using the facts that $\left(\mathbf{I}-\mathbf{H}_D\right)^{-1} = \mathbf{I} + \left(\mathbf{I}-\mathbf{H}_D\right)^{-1}\mathbf{H}_D$ and $\mathbf{W}_D\mathbf{H}_D = \mathbf{H}_D^T\mathbf{W}_D$, we have

$$
\begin{aligned}
MDFFIT_D &= \mathbf{e}_D^T \mathbf{W}_D \mathbf{H}_D \left(\mathbf{I} - \mathbf{H}_D\right)^{-1} \mathbf{e}_D \\
&= \mathbf{e}_D^T \mathbf{W}_D \mathbf{H}_D \left[\mathbf{I} + \left(\mathbf{I} - \mathbf{H}_D\right)^{-1} \mathbf{H}_D\right] \mathbf{e}_D \\
&= \mathbf{e}_D^T \mathbf{W}_D \mathbf{X}_D \left[\mathbf{A}^{-1} + \mathbf{A}^{-1} \mathbf{X}_D \mathbf{W}_D \left(\mathbf{I} - \mathbf{H}_D^T\right)^{-1} \mathbf{X}_D^T \mathbf{A}^{-1}\right] \mathbf{X}_D^T \mathbf{W}_D \mathbf{e}_D \\
&= \mathbf{e}_D^T \mathbf{W}_D \mathbf{X}_D \left(\mathbf{X}_{(D)}^T \mathbf{W}_{(D)} \mathbf{X}_{(D)}\right)^{-1} \mathbf{X}_D^T \mathbf{W}_D \mathbf{e}_D .
\end{aligned}
$$

The formulation of the deletion set *D* remains a problem when we try to apply the multiple-row deletion approach to the survey data. In this study we suggest the construction of the deletion set should depend on the sample design. If the sample is collected from a single stage sampling design and of a moderate sample size, a stepwise approach like the one proposed by Belsley, Kuh, and Welsch (1980) can be used to filter influential groups. If, on the other hand, the sample comes from a multi-stage complex design and is very large, we can use sampled PSUs or some specific characteristic group as the possible deletion sets or conduct a "forward" searching process. We will address this in next two sections.

4.2 Deletion of Specific Characteristic Groups

Large sample size is usually a feature of survey data, which will naturally cause computational difficulties in the process of influence analysis. The deletion groups discussed in Section 4.1 can be linked to characteristics of individuals, such as gender, race and age. In a household survey, units from certain demographic groups may be influential when their $\mathbf{Y}$ values, $\mathbf{X}$ values, or weights are distinct from those of other groups.

In some surveys, entire PSUs of units may be candidates for deletion. Consider a

household interview survey in which two PSUs are sampled per stratum. As noted above, if one data collector does all interviewing in a PSU, data for all sample units in the PSU may be affected. If one PSU is deleted, it can be treated as deleting a unit in a single-stage sample while each PSU is equivalent to an individual unit. The diagnostic statistics for single-stage samples, which were described in Chapter 3, are also suitable in this case, but the cutoff values should be related to the total number of PSUs, but not the total number of observations in the sample. By cycling through all sample PSUs, a set of group-deletion diagnostics can be obtained for judging the influence of individual PSUs. If some specific characteristic groups are suspected of being influential and they are across the PSUs, we may have to determine the cutoffs by intuition and empirical judgment, instead of directly borrowing cutoffs from the single-case deletion methods.

4.3 Forward Search

4.3.1 Introduction

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. Unlike the backward search, which applies the single-case deletion diagnostics repeatedly and therefore suffers from the combinatorial explosion of the number of cases, the forward

search starts by fitting a model using the robust method of least median of squares (LMS). The initial subset, recommended by Rousseeuw (1984), is determined to be the one, among a large number of randomly chosen subsamples of size $m = p$, where p is the number of regressors, yielding the parameter estimate $\mathbf{b}$ which minimizes the median of the squared residuals $e_i^2(\mathbf{b})$. The squared residuals are therefore calculated for all n observations in the original sample using $\mathbf{b}$ and ordered. The $m+1$ units with the smallest squared residuals are chosen to be the new larger subset. The search repeats and the values of $\mathbf{b}$ are recorded at each step. In the absence of the outliers, the parameter estimates and the plots of scaled residuals are likely to be stable and smooth. If there are outliers, they will enter at the end of the search often causing noticeable jumps in s^2, but not necessarily in $\mathbf{b}$. The core feature of the forward search is that masked outliers are not included in the initial subset. The least median of squares criterion can be replaced by that of least trimmed squares (LTS), which minimizes the sum of the smallest n^* squared residuals for some n^* with $[(n+p+1)/2] \le n^* \le n$. LTS estimates have a faster rate of convergence when the sample size is very large.

Since LMS procedures on estimate of $\boldsymbol{\beta}$ are unaffected by sample outliers, an obvious question is: why not simply use LMS on the full sample for model fitting? For one thing LMS does not identify particular observations as influencing the regression fit. Thus, that detailed information would be lost to an analyst. Also, LMS has not been adapted for use with survey weights and, thus, has no obvious design-based interpretation. Modifying LMS to fit more into design-based analysis could be a future research topic.

4.3.2 Adaptation to Survey Data

The forward search method is intriguing for survey data analysis because of its robustness for identifying a group of outliers and its computational feasibility. In this research we modify Atkinson and Riani's method and make it implementable to survey data. Before modification, the method allows observations other than the masked outliers to enter and leave the subset used for model fitting, uses the squared residuals as the filtering criterion, and tracks the mean squared error s^2 for outlier monitoring. For complex survey data, s^2 may not be reasonably estimable if the underlying model deviates from (2.1). Therefore, we consider other statistics for filtering and monitoring outliers. Here is a general description of how the forward search method may be modified and implemented in a single-stage 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 the key statistic which measures the change in regression parameters if this observation were removed from the subset.

(3) Retain the observation with the minimum key statistic, or in other words, retain the observation which causes the smallest change in regression if it were removed from the subset of size $m+1$.

(4) Repeat steps (2)-(3) until all observations are included in the regression.

By tracking the values of the parameter estimates and the key statistics, this algorithm should identify the point or points most influential in the model fitting. As

the algorithm proceeds sequentially through the points, outlying values will enter last. Therefore, the key statistic is expected to indicate abrupt changes in parameter estimates when the outliers begin to be introduced into the regression.

There are three important issues for this algorithm to function appropriately. The first is the choice of the initial subset. The initial subset should be free of outliers and have 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. We either keep the points among those having the smaller leverages, residuals, DFBETAS, DFFITS, and modified Cook's Distance, or keep a group with the minimum median of squared residuals (LMS). Both the single-case deletion diagnostic statistics and the LMS algorithm can be helpful for finding an outlier-free initial subset. Choosing the key statistics is the second important issue. During the forward searching process, the key statistics are used to monitor the changes in the regression while new observations come into the subset. Diagnostic statistics based on single-case deletion are possible candidates for the key statistics, among which modified Cook's Distance is more suitable because it summarizes the changes in all regression parameters and has stable performance. Other statistics, including the multiple-case deletion versions of DFBETA and DFFIT, can also be tracked to facilitate the judgment. The third issue is to draw a line between the outliers and the non-outliers. The cutoff value for the key statistic remains a problem in the forward search process. An analyst may simply use a fixed cutoff, such as 2 or 3, developed in Chapter 3. However, we suggest making a case-by-case judgment in which the analyst can account for the changing trends of both the key statistic and other available statistics.

A line may be drawn at the point after which the monitoring statistics have abrupt increases.

Once again, it is crucial to emphasize the importance of starting the searching process with a subset free of outliers in the modified forward search method. The method should not be sensitive to the choice of initial subset, provided outliers are not included at the start. Hence, we recommend that different initial subsets and various key statistics be applied to complete multiple searching processes so that we can confirm that 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. For example, at least 2 units need to be selected from at least one sample PSU in each sample stratum. If only one PSU is represented from a stratum in the initial set, special variance estimation procedures will be needed as described in Wolter (1985). For that reason it will typically be more convenient to select units for the initial set from two or more PSUs per stratum, assuming that design has multiple PSUs in each stratum.

Chapter 5: Application of Diagnostic Techniques for Influence Analysis

5.1 Introduction

This Chapter will document the performance of the proposed and modified statistics in Chapter 3 and Chapter 4. In order to verify and justify the effectiveness of these statistics on identifying influential observations, a logical approach is to apply them to real survey data and then conduct appropriate evaluations. I will employ two survey data sets in Section 5.2 and 5.3: the 1998 Survey of Mental Health Organizations (SMHO) and the 1999-2002 National Health and Nutrition Examination Survey (NHANES). Both of the surveys contain a variety of variables that are suitable for linear regression analysis.

The 1998 SMHO collected data on approximately 1,530 specialty mental health care organizations and general hospital mental health care services, with an objective to develop national and state level estimates for total expenditure, full time equivalent staff, bed count, and total caseload by type of organization. The universe of mental health care organizations not only includes large sample units such as the state and county mental health hospitals, private psychiatric hospitals, multi-service mental health organizations, Department of Veteran Affairs medical centers, and nonfederal government hospitals with separate psychiatric services, but also includes some small units such as residential treatment centers, free standing outpatient clinics, and partial-care organizations. The sample for this survey was based on a stratified single-stage design with probability proportional to size (PPS) sampling. The primary strata were defined on the basis of type of organization, ownership type, and type of setting. The varying sizes of the mental health care organizations result in the values of collected variables in the sample having wide ranges, which may cause some observations to have relatively large influence on the parameter estimates of a linear regression.

The NHANES survey is conducted by the National Center for Health Statistics, Centers for Disease Control. There are several of these data sets publicly available,

including the most recent ones, 1999-2000, 2001-2002, and 2003-2004. This survey is a rich source of quantitative and qualitative variables which are designed to assess the health and nutritional status of adults and children in the United States through interviews and direct physical examinations. NHANES uses a complex, multistage, probability sampling design. Oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. The data set used in our study is a subset of 1999-2002 data composed of Mexican-American women aged 20-29. Due to oversampling of Mexican-Americans, the final weights in our sample range from 698.39 to 103,831.17. This is a ratio of 149:1 for the largest weight to the smallest.

The two surveys have different design features and variables with different properties. Therefore, two case studies will be conducted in this Chapter, using SMHO and NHANES data. While both case studies will examine the performance of the single-case deletion statistics proposed and modified in Chapter 3, they emphasize different survey designs and adopt different variance estimation methods and cutoff values. Section 5.2 will present the first case study using SMHO data, whereas the results from the second case study using NHANES data will be demonstrated in Section 5.3. In Section 5.4 simulations are used to study the performance of the diagnostics in a more controlled setting. A pseudo population will be constructed from SMHO data, based on which we will evaluate and compare the application of single-case deletion techniques to the regression estimation. In Section 5.5 and Section 5.6 we will revisit the two case studies and the simulation by employing the forward search method which is designed to identify influential groups. Computational and graphical work in this dissertation was mainly done by R software.

5.2 Identifying Single Influential Observations: Case Study 1

5.2.1 Summary of SMHO Data Set

The model of interest in this study is to regress the total expenditure of a health organization on the number of beds set up and staffed for use and the number of additions of patients or clients during the reporting year. The total expenditure was defined as the

sum of salary and contract personnel expenses, other contract and operating expenses, and depreciation expenses, and then divided by 1000. The number of beds accounted for hospital bed count and residential bed count. Similarly, the number of additions included hospital additions count and residential additions count. Scatterplots of expenditures versus beds and additions are shown in Figure 5.4 later in this Chapter. We ignored the stratification and substratification in the sampling design and treated the sample as selected from a single-stage sampling with varying selection probabilities. The effect of stratification and clustering on the variance estimation and the statistics used to identify influential observations will be addressed in the second case study. A total of 875 observations were used in the above regression due to missing values in the independent and dependent variables.

Table 5.1 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 519,863.27, which is almost 30,000 times the minimum, 16.6. Although not as tremendous 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 from a PPS design, the sample weights were associated with the sizes of the mental health organizations, with a range from 0.99 to 158.86. Since the ranges of expenditures, beds, and additions are large, an option would be to transform, e.g., by taking logs, before fitting a model. We have not pursued that here.

Table 5.1. Quantiles of Variables in SMHO Regression.

	Quantiles				
Variables	0%	25%	50%	75%	100%
Expenditure (1000's)	16.6	2,932.5	6,240.5	11,842.6	519,863.3
# of Beds	0	6.5	36	93	2405
# of Additions	0	558.5	1410	2406	79808
Weights	0.99	1.42	2.48	7.76	158.86

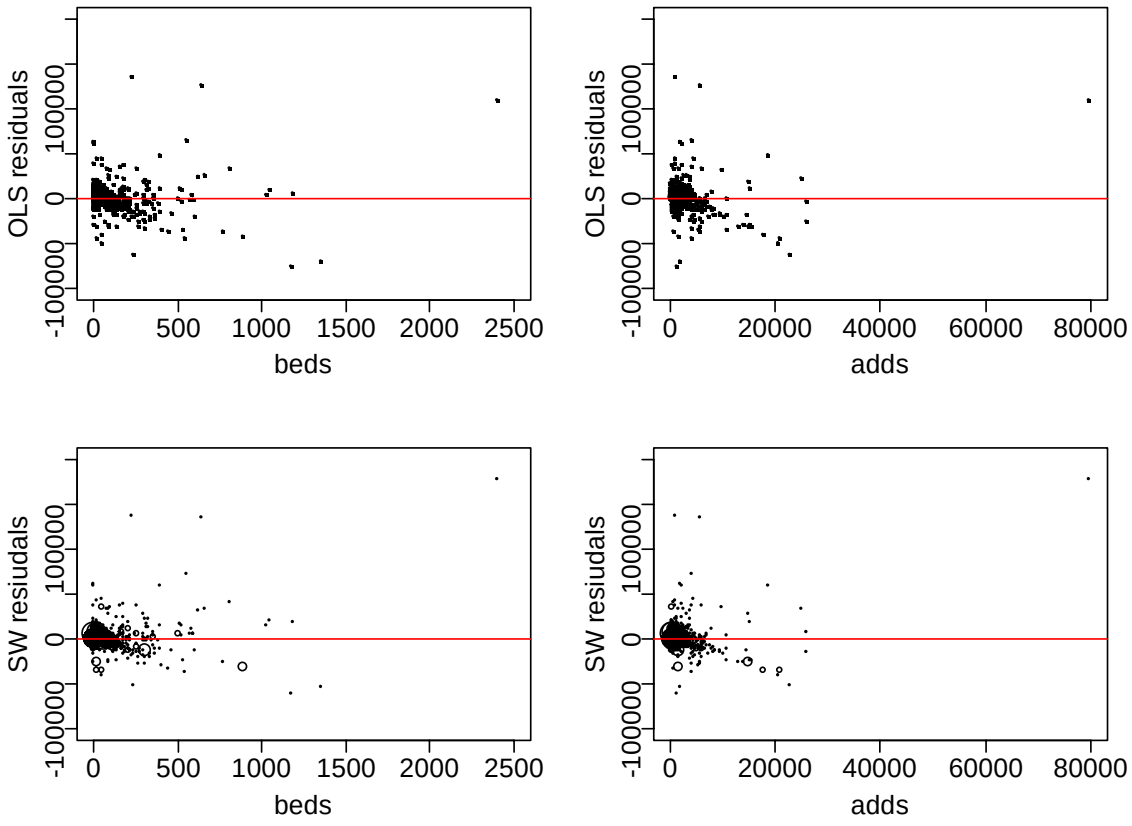
5.2.2 Parameter Estimation

The identification of single influential points will be compared under two different settings. One is to assume the sample is selected from a simple random sampling (SRS) design and analyzed by conventional OLS regression estimators. This approach might be used by an analyst who elected to ignore all design features. The other is to assume a single-stage sampling with varying sample weights which will be incorporated into the regression estimation. The estimated coefficients and their standard errors are reported in Table 5.2, with SW denoting survey weighted estimates. The intercept and slope coefficients all have discrepancies between the two methods, and the estimated intercept even changes from negative to positive and from significant to insignificant. The relative size of the differences between the OLS and SW estimates is much greater for the intercept than the slopes. Analysts are often more focused on the slope estimates. The effect of survey weights on coefficient estimation signals that survey weights could play a crucial role in influence analysis on this regression. Figure 5.1 shows scatterplots of the OLS and SW residuals versus the two auxiliary variables. Bubble plots were drawn for the SW regressions, in which areas of the bubbles are proportional to the sizes of sample weights.A few observations with extreme $\mathbf{X}$ values also have large residuals and therefore could be possible influential units that greatly affect the parameter estimates. The OLS and SW residuals have similar patterns but the values can be quite different, for instance, the SW residual of the point in the upper right corner of each scatterplot is larger than the corresponding OLS residual.

Table 5.2. OLS and SW Parameter Estimates of SMHO Regression of Expenditures on Beds and Additions.

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	-1201.73	526.19	-2.28	514.08	1157.71	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

Figure 5.1. OLS and SW residuals versus Two Auxiliary Variables for SMHO Data. The red lines were drawn at residuals equal to zero.



From the next section on, the results of applying the diagnostic approaches will be displayed in tables and plots. In the plots reference lines will be drawn at the cutoff values where appropriate. For the SW diagnostics, a loose criterion, 3, was used to construct cutoffs. For example, the cutoff of DFBETAS is $\frac{3}{\sqrt{n}}$, and the cutoff of DFFITS is $3\sqrt{\frac{p}{n}}$. However, dotted lines will also be drawn at the cutoff values constructed on the stricter criterion, 2. As in Figure 5.1, bubble plots were drawn for the SW regressions and diagnostics.

5.2.3 Diagnostics by Leverages and Residuals

Figure 5.2, on the left, shows a scatterplot of leverages calculated using two methods with and without sample weights. Outlying points, with leverages greater than twice

their mean, were identified to be the ones beyond the two reference lines. The 27 outlying observations identified by the SW but not by the OLS diagnostics, represented by relatively large bubbles in area A, are associated with large sample weights ranging from 7.44 to 158.86; whereas the 14 outlying observations identified by the OLS only, represented by small bubbles in area B, have small weights ranging from 0.99 to 2.62. The bubbles in the upper right square, with moderate sizes, stand for the points identified by both methods. The small dot in the upper right corner is an observation with extremely large total expenditure, number of beds, number of additions, but a small sample weight. Later we will show that it is always associated with large diagnostic statistics.

The points in the residual plot on the right show the residuals scaled by the estimated standard error $\hat{\sigma}$ of model (2.1), where $\hat{\sigma}$ was estimated by the OLS estimator for the OLS scaled residuals and by the SW formula (3.3) for the SW scaled residuals. With a few exceptions, the weighted and unweighted diagnostics identified similar extreme residuals. The residual analysis mainly filters out the observations with outlying Y values, but not necessarily those with outlying weights.

Figure 5.2. Leverage and Residual Diagnostic Plots for SMHO Data. In the leverage plot on the left, area A includes points identified as outlying by the SW diagnostic only, whereas area B includes points identified by the OLS diagnostic only. In the residual plot on the right, areas A and B include points identified by SW only, whereas areas C and D include points identified by OLS only. The red line was drawn at 45 degrees.

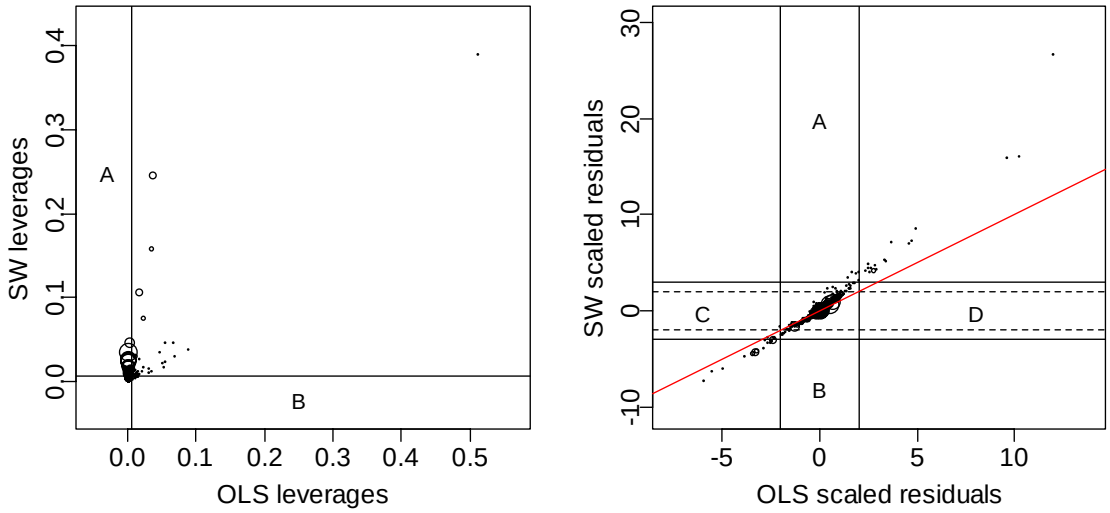


Table 5.3 and Table 5.4 list the coefficient estimates on the reduced samples excluding the identified outlying observations. Qualitatively, the conclusion would be the same whether one uses OLS or SW diagnostics – all three parameter estimates are significantly different from zero. A more quantitative measure of difference is obtained by comparing predicted values calculated after excluding units identified by the OLS and SW diagnostics. Figure 5.3 displays the resultant fitted values versus those from the full samples. The slope coefficients decreased when the outliers were not used in the regressions, which accordingly resulted in smaller fitted values. In Figure 5.3 we observe that some outliers tend to be associated with larger changes in the prediction of $\mathbf{Y}$ between including and excluding them in the sample. The OLS and the SW parameter estimates from the reduced samples may be quite different from each other, as we can see in Table 5.3 and 5.4. As a result, the OLS and the SW fitted values computed using those estimated parameters can also be far apart. In the two scatterplots in the third row of Figure 5.3, it is shown that by applying leverage diagnostics the SW estimator produced larger slope estimates and therefore larger fitted values, whereas the OLS estimator yielded relatively big predictions in expenditures when residuals were

used to identify outliers.

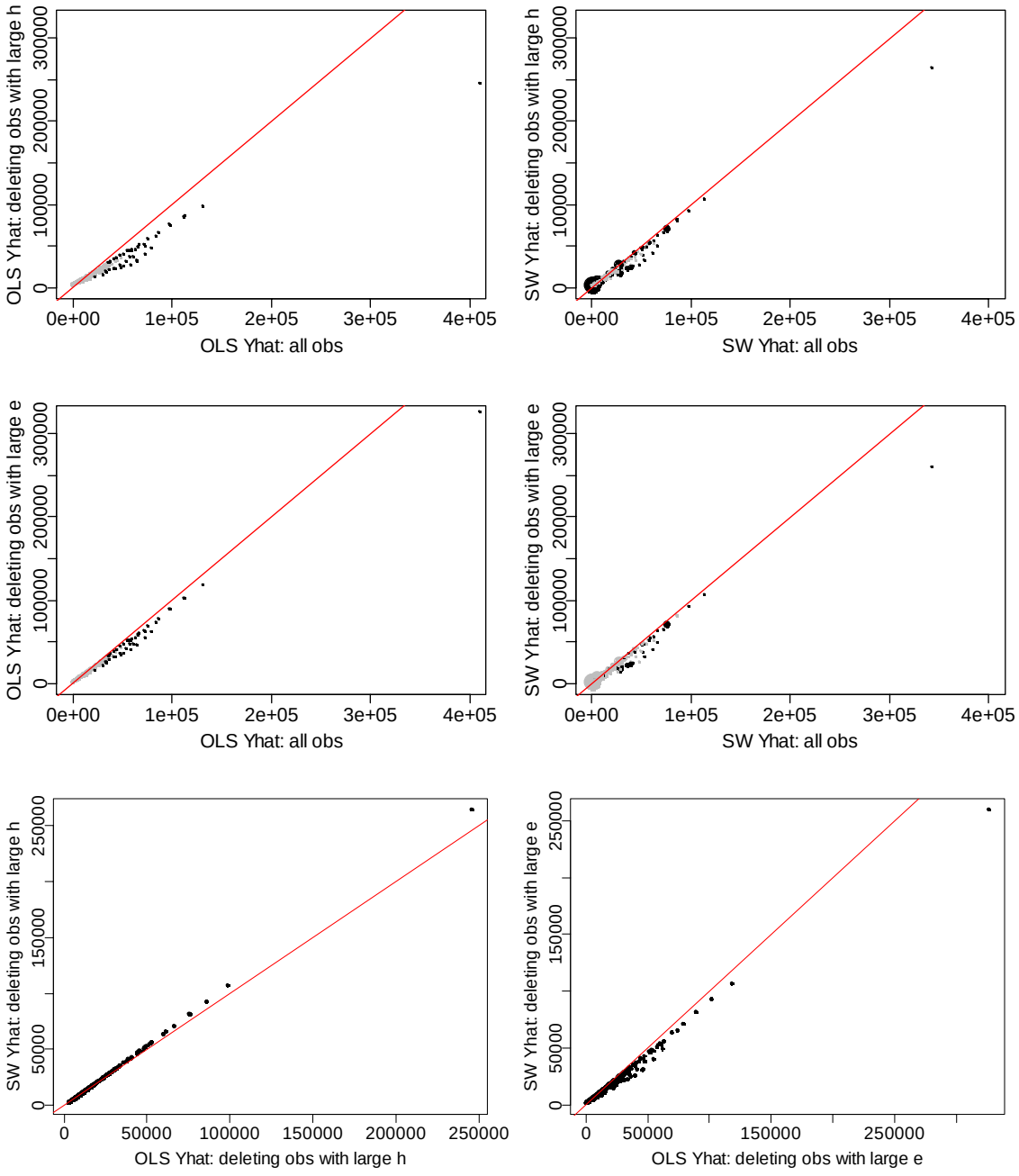
Table 5.3. OLS and SW Parameter Estimates after Deleting Observations with Large Leverages from SMHO Regression.

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	2987.55	490.54	6.09	1993.86	353.71	5.64
# of Beds	69.27	4.347	15.94	75.82	6.75	11.23
# of Additions	0.947	0.201	4.71	0.997	0.211	4.73

Table 5.4. OLS and SW Parameter Estimates after Deleting Observations with Large Residuals from SMHO Regression.

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	645.83	311.63	2.07	1674.66	386.27	4.34
# of Beds	84.48	1.98	42.67	76.19	5.28	14.43
# of Additions	1.531	0.103	14.86	0.932	0.217	4.29

Figure 5.3. Fitted Values Plots After Applying Leverage and Residual Diagnostics to SMHO Data. In the first two rows are the fitted values from the regression on sample deleting observations with large leverages or large residuals versus those from the regression on full sample, both OLS and SW. Points in grey are ones not identified by the diagnostics; points in black are ones identified as influential. In the third row are the OLS fitted values versus the SW fitted values from the regression on sample excluding outliers identified by OLS and SW. A 45 degrees line is drawn in each panel.



5.2.4 Diagnostics by DFBETAS

The diagnostic results of the DFBETAS statistics for number of beds and number of additions are graphically presented in Figure 5.4. It conveys similar messages as the leverage diagnostics in Figure 5.2. It is clearly shown, especially in the partially enlarged graphs at the second row, that points identified only by the OLS method have small weights symbolized by the bubbles of small sizes. Using the SW formula of DFBETAS, we singled out a few points associated with moderate sampling weights though almost all of them were also identified by OLS. Figure 5.5 includes scatterplots of total expenditure versus the two auxiliary variables, which indicate the positions of the identified cases relative to the scatterplot smoothing lines. In the OLS case this line was fitted using the *lowess* function in R STATS package, whereas in the SW case it was done by the *svysmooth* function in R SURVEY package. It is worth attention that there is an extremely outlying point located at the upper right corner of each graph. This point corresponds to the hospital with the largest number of beds and additions and largest value of expenditure in the example. We expect the parameter estimates are likely to become smaller, to different extents, if this point were eliminated.

Figure 5.4. DFBETAS Plots for SMHO Data. Areas A and B include points identified only by the SW diagnostics whereas areas C and D include points identified by the OLS diagnostics only. The partially enlarged graphs are presented below the originals.

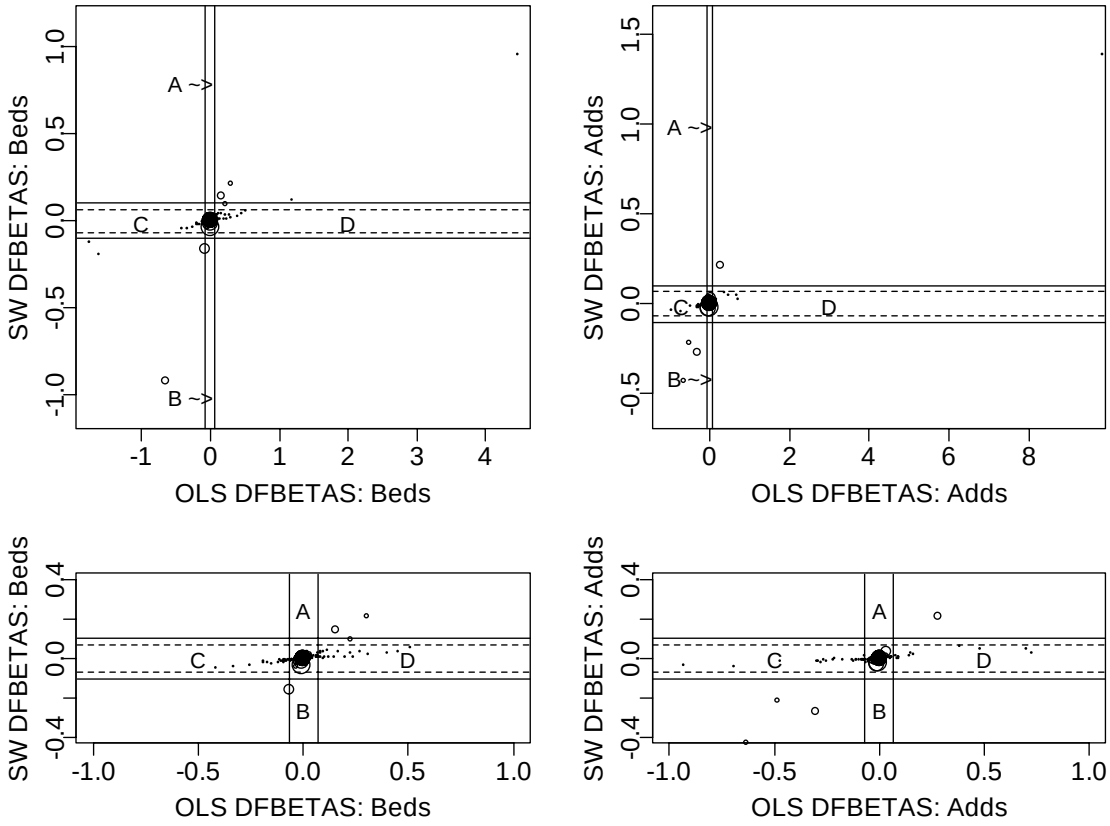
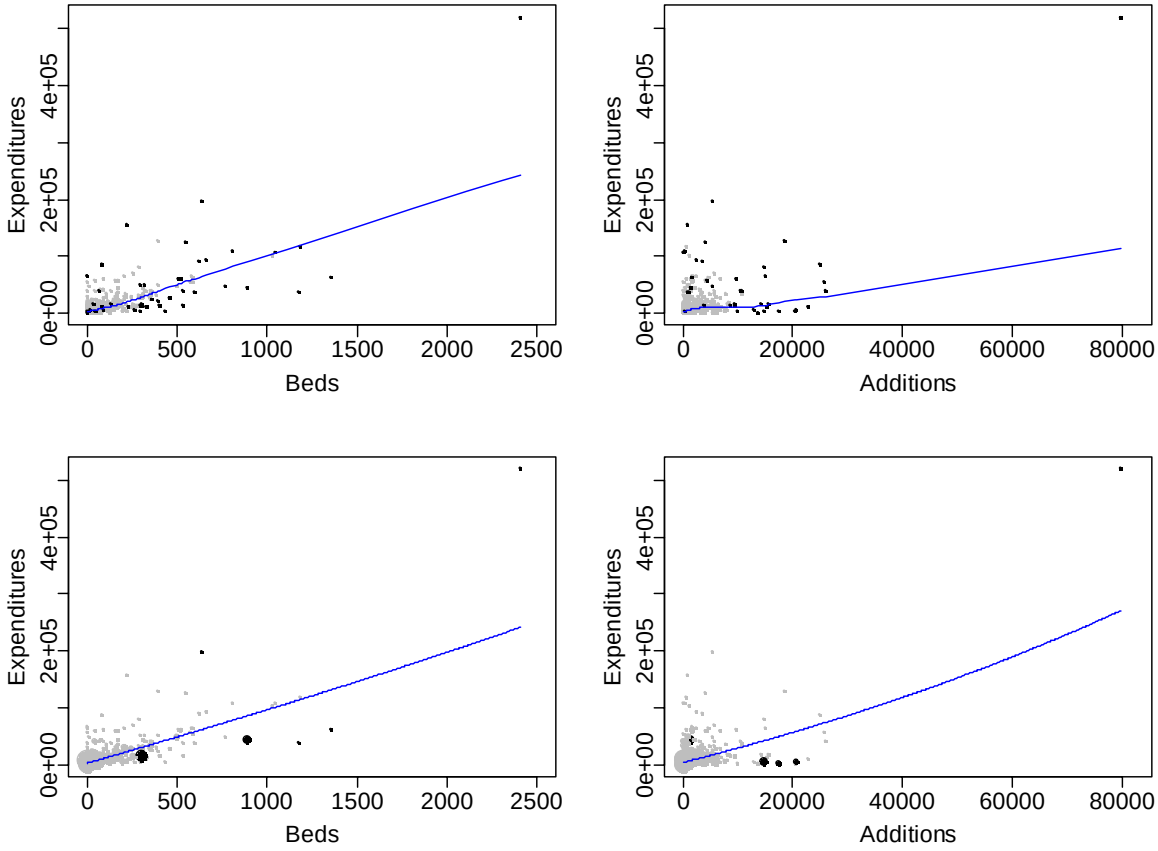
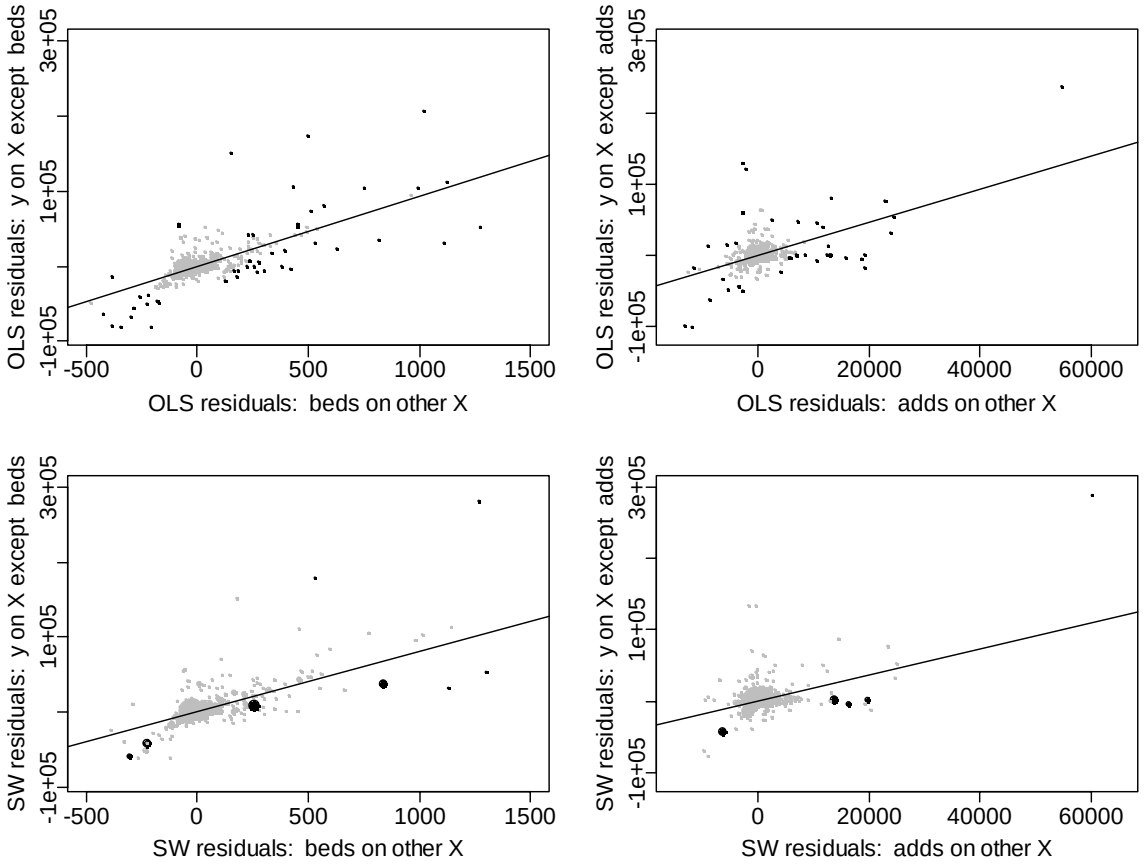


Figure 5.5. Scatterplots with OLS (top) and SW (bottom) Smoothing for SMHO Data. The dark dots symbolize the points identified as influential by the OLS or SW DFBETAS statistics.



Another way to show how the deletion of an observation affects each coefficient estimate is to draw an added variable plot, as we introduced in Section 3.6.1. Figure 5.6 displays two sets of added variables for two auxiliary variables and for the OLS and the SW regressions, respectively. In the OLS plots, the identified influential points labeled as dark are scattered around the corners where they deviate further from the middle of the regression line than the unidentified points. However, in the SW plots, the dark dots are not necessarily the furthest away from the center of the regression line if they are associated with very large sampling weights. There are even some points which stray greatly from the rest but are not identified because their weights are too small.

Figure 5.6. OLS and SW Added Variable Plots for SMHO Data. The dark dots indicate the influential observations identified by OLS and SW DFBETAS statistics. The lines are OLS (top row) regression fits or WLS (bottom row) regression fits which have the same slope as the parameter estimate for beds (or additions) in full sample.



Tables 5.5 through Table 5.7 report the estimated coefficients and their standard errors when the identified outliers were removed from the sample. Excluding the observations with large DFBETAS for number of beds, we obtained a slightly larger SW slope estimate for number of beds, meanwhile the estimated slope for number of additions greatly dropped to 1.03. For the OLS estimates, both slopes moderately decreased. Hence, the OLS fitted values became smaller but the SW ones were less affected (See graphs at the first row of Figure 5.7). When deleting the cases with large DFBETAS of number of additions, the parameter estimates declined for both OLS and SW, but the OLS estimates have larger changes. The estimated slope of number of additions even dropped from 2.31 to 0.79, which resulted in smaller fitted values in the OLS graph at the second row of Figure 5.7. Still, the fitted values from the SW regression only changed to a small extent. Table 5.7 and the last two graphs in Figure

5.7 show the regression estimates and the fitted values after deleting the observations with either large DFBETAS of number of beds or large DFBETAS of number of additions. The estimates were in between the results from deleting only one kind of outliers. Note that the SW SEs in Tables 5.5, 5.6, and 5.7 are substantially smaller than the SEs in Table 5.2 where all points were used. This is expected because the points that are deleted are much different from those that are retained.

Table 5.5. OLS and SW Parameter Estimates after Deleting Observations with Large DFBETAS of Beds for SMHO Data.

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	704.97	364.98	1.93	1654.12	436.53	3.79
# of Beds	83.06	2.91	28.54	82.73	4.53	18.26
# of Additions	1.841	0.128	14.38	1.034	0.321	3.22

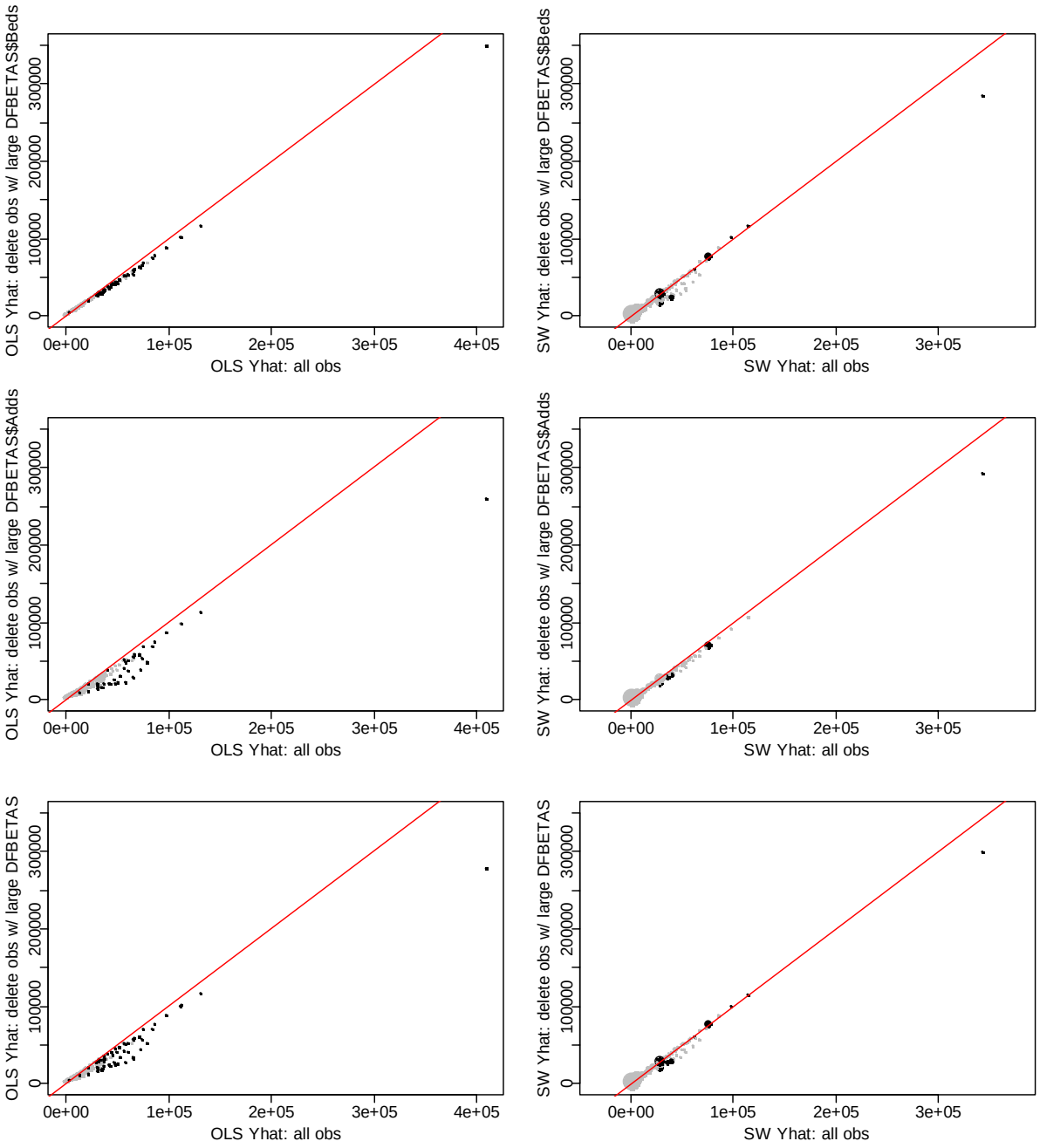
Table 5.6. OLS and SW Parameter Estimates after Deleting Observations with Large DFBETAS of Adds for SMHO Data.

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	2463.11	403.57	6.10	1565.4	444.39	3.52
# of Beds	80.47	2.54	31.68	75	6.61	11.34
# of Additions	0.79	0.17	4.65	1.382	0.275	5.03

Table 5.7. OLS and SW Parameter Estimates after Deleting Observations with Large DFBETAS of either Beds or Adds for SMHO Data.

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	2044.54	353.01	5.79	1485.03	425.83	3.49
# of Beds	82.36	2.61	31.55	81.72	4.49	18.19
# of Additions	0.96	0.15	6.42	1.27	0.28	4.59

Figure 5.7. Fitted Values Plots After Applying DFBETAS Diagnostics to SMHO Data. The OLS and SW fitted values are from regressions on sample deleting observations with large DFBETAS for beds, DFBETAS for additions, or either. The red lines are drawn at 45 degrees.



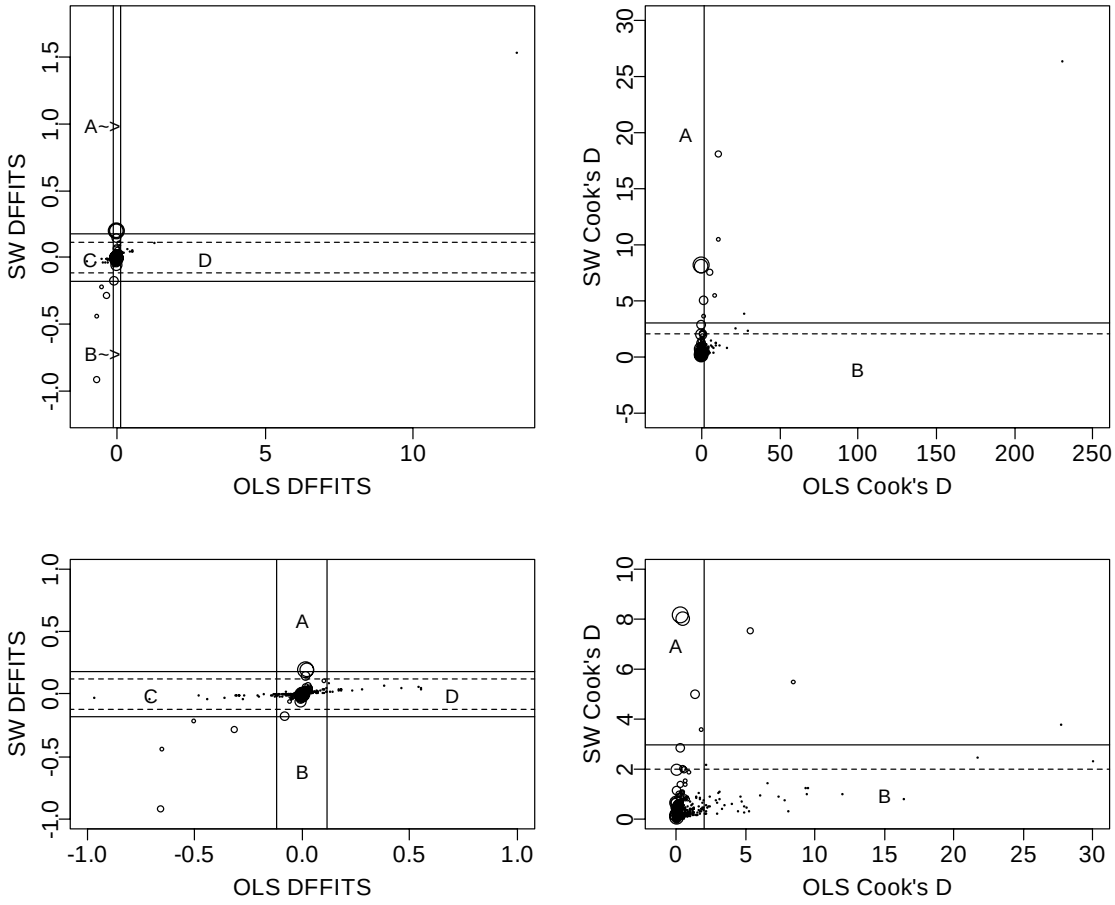
The OLS diagnostics identified too many points as being influential compared to the SW diagnostics. This led to systematic reductions in predicted values for OLS predictions when these points were omitted. The SW analysis omits fewer points and

has less of an effect on predictions. Thus, if an analyst takes the position that the sample design is ignorable, does not use weights, and applies OLS diagnostics, substantially different predictions would be obtained in this case.

5.2.5 Diagnostics by DFFITS and Modified Cook's Distance

Both DFFITS and modified Cook's Distance statistics summarize the effect of deleting a specific unit on the overall parameter estimation. There were 3 influential observations identified by the SW DFFITS but not by the OLS diagnostics in Figure 5.8, with their weights ranging from 37.8 to 158.86. There are 39 influential observations identified by the OLS DFFITS only. Their weights were relatively small, ranging from 0.99, which is the smallest weight in the sample, to 5.5. The SW modified Cook's Distance exclusively identified 4 cases with weights from 11.38 to 158.86, whereas the OLS Cook's Distance only uniquely detected 38 points with weights that range from 0.99 to 5.5. None of the cases with large weights were identified by the OLS Cook's Distance.

Figure 5.8. DFFITS and Modified Cook's Distance Plots for SMHO Data. Areas A and B in the DFFITS plot and area A in the Cook's Distance plot include points identified only by the SW diagnostics, whereas areas C and D in the DFFITS plot and area B in the Cook's Distance plot include points identified by the OLS diagnostics only. The partially enlarged graphs are presented below the originals.



There was only one observation identified by the OLS modified Cook's Distance but not by the OLS DFFITS. Therefore, the parameter estimates based on the samples without the identified outliers are very similar for these two cases. The estimated slopes dropped moderately compared to the ones from full sample, which correspondingly caused smaller fitted values. Most of the outliers are associated with relatively large changes in fitted values. For the SW diagnostics, the two statistics also have comparable performance. Since fewer outliers were picked from the sample by the SW DFFITS and the SW modified Cook's Distance, Table 5.8 and Table 5.9 illustrate that the SW estimates from the reduced samples changed less than the OLS ones. Comparing to Table 5.2, we see that the SEs again decrease substantially after deleting cases,

particularly for SW. Figure 5.9 shows that the fitted values did not deviate very much from those on the full sample when the SW diagnostics are used to determine which points to eliminate.

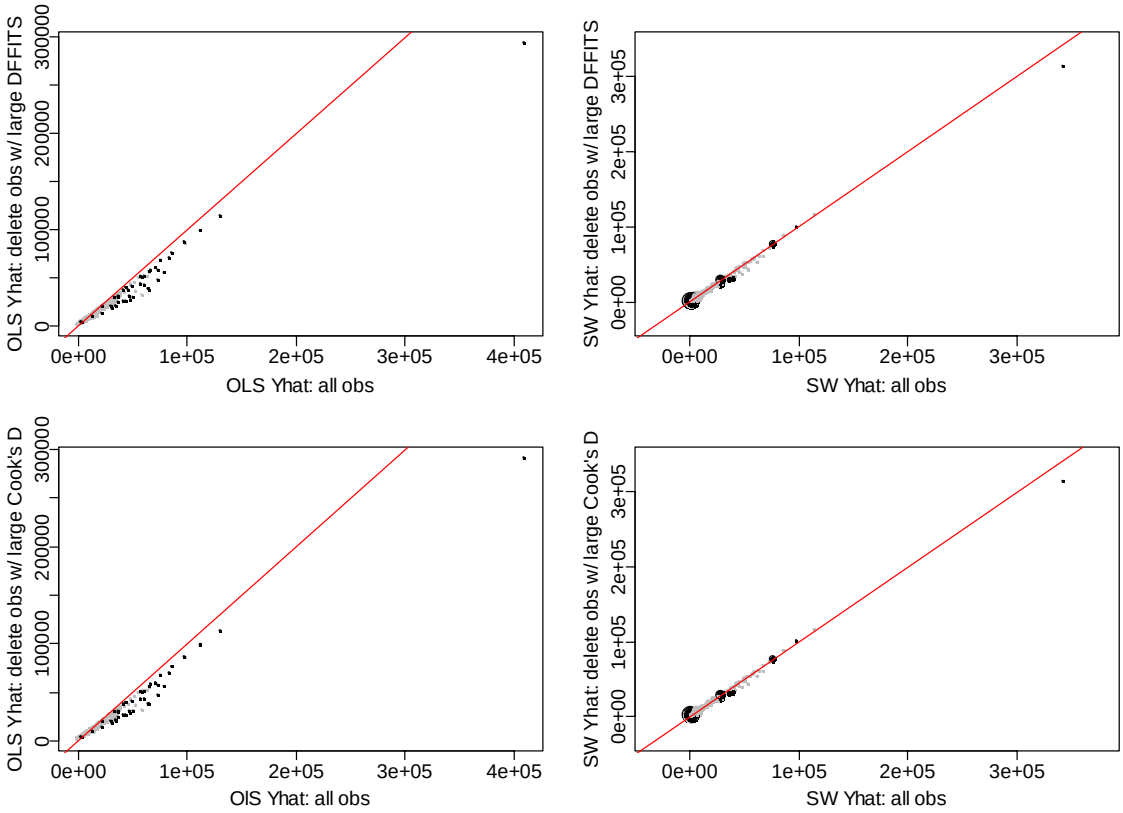
Table 5.8. OLS and SW Parameter Estimates after Deleting Observations with Large DFFITS for SMHO Data.

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	1617.67	335.38	4.82	1028.71	360.46	2.85
# of Beds	81.45	2.44	33.38	82.94	5.72	14.50
# of Additions	1.20	0.12	9.77	1.40	0.27	5.27

Table 5.9. OLS and SW Parameter Estimates after Deleting Observations with Large Modified Cook's Distance for SMHO Data.

Independent	OLS Estimation			SW Estimation		
Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	1660.45	335.54	4.95	932.43	345.86	2.70
# of Beds	80.92	2.44	33.16	82.83	5.72	14.48
# of Additions	1.19	0.12	9.66	1.43	0.26	5.43

Figure 5.9. Fitted Values Plots After Applying DFFITS and Cook's D Diagnostics to SMHO Data. The OLS and SW fitted values are from regressions on sample deleting observations with large DFFITS and Modified Cook's Distance. The red lines are drawn at 45 degrees.



5.2.6 Discussion

The conventional OLS influence diagnostics were adapted in previous chapters to be used for survey data. The cutoff values for the adapted statistics were 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 the SMHO sample, we conclude that the SW diagnostics, including leverages, residuals, DFBETAS, DFFITS, and modified Cook's Distance, identify different points than the OLS diagnostics as being influential. This is because in the SW regressions, points can be influential due to outlying sample weights besides extreme **Y** and **X** values. Different diagnostic approaches identify different sets of influential observations because they focus on measuring diverse kinds of changes in the regression estimation after a point is deleted from the sample. Therefore, a researcher should apply

appropriate diagnostic statistics to the analysis depending on what types of outliers he intends to detect.

Note that there can be situations where points with large weights, residuals, or $\mathbf{X}$ values would be important in identifying whether a model is correctly specified. For example, if Y were quadratically related to an $\mathbf{x}$ and units with large $\mathbf{X}$'s were deleted because of large weights or large residuals, the ability could be lost to recognize that the model should be quadratic. Thus, the diagnostics studied here should be applied with care.

5.3 Identifying Single Influential Observations: Case Study 2

5.3.1 Summary of NHANES Data Set

In the second case study we examined a regression of systolic blood pressure on the logarithm of blood lead level, age, and body mass index using a subset from NHANES 1999-2002. A similar linear regression analysis has been done with a different sample by Korn and Graubard (1999), and the regression results are presented in Chapter 6 of their book. The subset used in this study has a sample size of 810, consisting of Mexican-American females aged 20 to 29. Unlike Case Study 1, this sample does not have very skewed $\mathbf{Y}$ and $\mathbf{X}$ values, but involves clustering and stratification in the sampling design with a set of large and greatly varying sample weights. There are $n = 57$ PSUs nested in $H = 28$ strata, most of the strata having 2 PSUs. The average cluster size $\bar{m}$ is 14.21 persons. When applied to a clustered data set, the variance estimators in the SW diagnostic statistics need to take the design into account and the cutoffs for some of the statistics contain an estimate of ρ, which in model (3.9) describes the correlation between the observations within the same cluster. The illustrative calculations in this study do not account for the fact that Mexican-American females are a domain within the full population. This will tend to make SW variance estimates smaller than they would be if the domain feature was accounted for.

Table 5.10 gives the quantile values of the variables and sample weights used in the regression. Besides demonstrating the skewness and large magnitude of sample weights,

it also shows that BMI and the logarithm of the blood lead are skewed to the right of their distributions, but the skewness is much smaller than that of the sample weights. Since the minimum of the originally measured blood lead level is as small as 1, we added 1 to blood lead level before took the logarithm to generate positive transformed values (Adding 1 is often done to avoid taking the log of zero; this step was not strictly necessary here). Note that using the untransformed value of blood lead would have resulted in more extreme **X** values. However, this type of modeling has previously been done using the log transformation (see, Korn and Graubard 1999), and we follow that precedent here. Figures 10 and 11 respectively display plots of systolic blood pressure and residuals versus the three auxiliary variables. Table 5.11 reports the parameter estimates of the regressions with and without weights. The SW estimators produced slightly larger intercept and slightly smaller slope of BMI than the OLS ones. Both methods agreed that age and blood lead do not have significant effects in determining the systolic blood pressure. Therefore, in the following diagnostic analysis, we will only focus on the changes in the estimated coefficient of BMI.

Table 5.10. Quantiles of Variables in NHANES Regression of Systolic Blood Pressure on Age, BMI, and Blood Lead.

	Quantiles				
Variables	0%	25%	50%	75%	100%
Systolic BP	82	102	108	114	146
Age	20	22	24	27	29
BMI	14.42	22.84	26.43	31.62	61.68
Log(Lead+1)	0.18	0.47	0.64	0.83	3.75
Weight	698.39	3576.69	11467.06	31094.18	103831.17

Table 5.11. OLS and SW Parameter Estimates from NHANES Regression.

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

*** significant at level 0.000

Figure 5.10. Bubble Plots of Systolic Blood Pressure versus Three Auxiliary Variables for NHANES Data. The sizes of bubbles are proportional to sample weights.

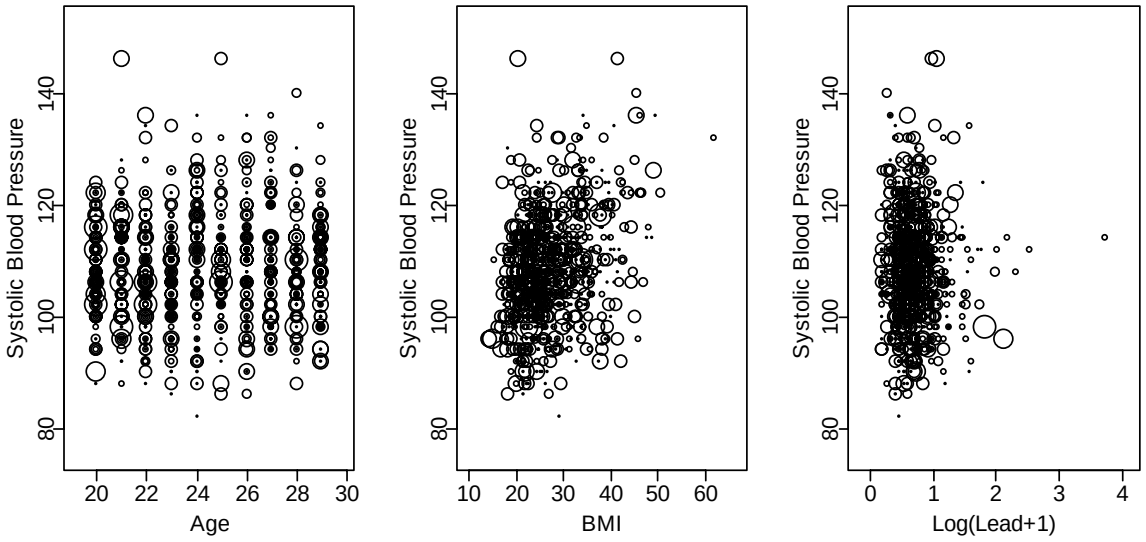
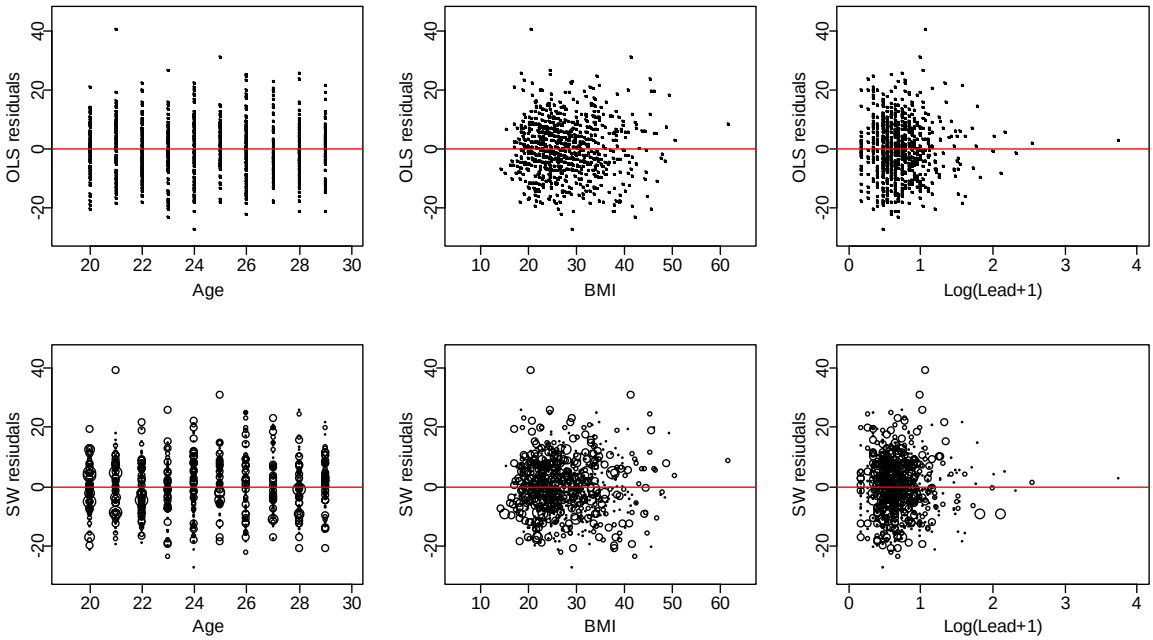


Figure 5.11. OLS and SW residuals versus Three Auxiliary Variables for NHANES Data. The red lines were drawn at residuals equal to zero.



5.3.2 Diagnostic Results

Similar to Case Study 1, we applied both the OLS and the SW diagnostic statistics, such as leverage, residuals, DFBETAS, DFFITS, and modified Cook's Distance to the regression estimation. Since the sample weights were not separately provided at cluster

level and at unit level, the parameters ρ and σ^2 in model (3.9) can only be estimated using the purely model based estimator in Section 3.5.3. Utilizing the VARCOMP procedure in SAS, we obtained $\hat{\rho} = 0.033$ and $\hat{\sigma}^2 = 82.09$. The design effect was estimated as $\sqrt{1+\hat{\rho}(\bar{m}-1)} = 1.2$. For the SW diagnostics, a strict criterion, 2, was used to construct cutoffs. For example, the cutoff of DFBETAS is $\frac{2}{\sqrt{n\bar{m}\left[1+\hat{\rho}(\bar{m}-1)\right]}}$. The solid reference lines in the subsequent figures were drawn at the cutoff values, and the dotted reference lines constructed using the loose criterion, 3, were also drawn in the same graphs.

Figure 5.12 through 5.14 display the comparisons between the OLS and the SW diagnostic statistics. The NHANES data set has widely-spread sample weights. Hence the SW diagnostics tend to identify more influential observations with large weights, whereas the OLS diagnostics tend to detect more points with small weights. The leverage plot, DFBETAS plot, and the modified Cook's Distance plot clearly show that the "identified by SW only" areas contain many big bubbles, but the "identified by OLS only" areas are filled with small dots. The residual plot is an exception in which the OLS and the SW residuals are very similar. This is mainly because the $\mathbf{Y}$ and $\mathbf{X}$ values in the data set are not extremely outlying.

Figure 5.12. Leverage and Residual Plots for NHANES Data.

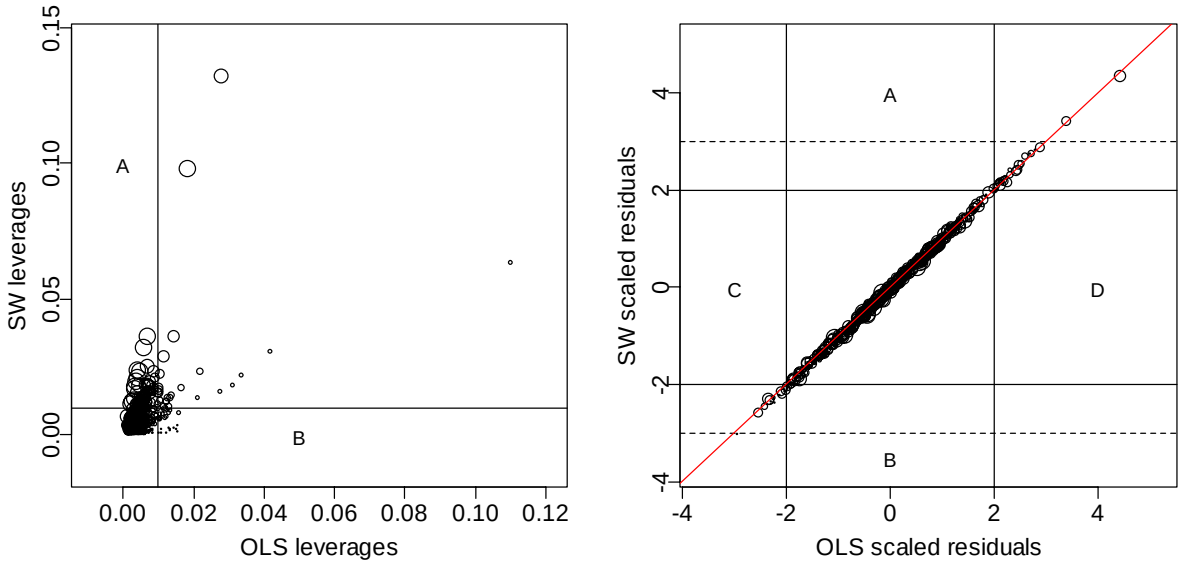


Figure 5.13. DFBETAS Plot and Added Variable Plots of BMI for NHANES Data.

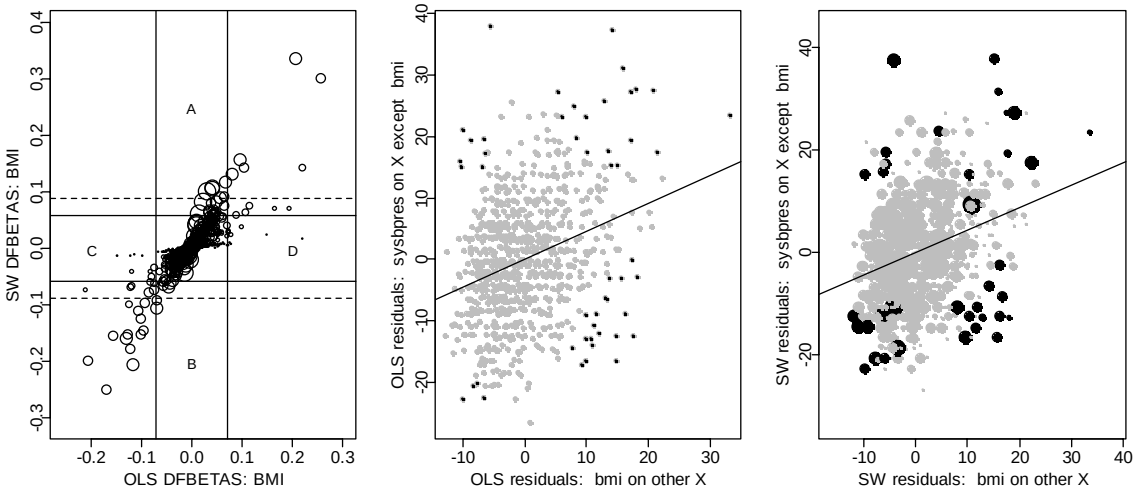


Figure 5.14. DFFITS Plot and Modified Cook's Distance Plot for NHANES Data.

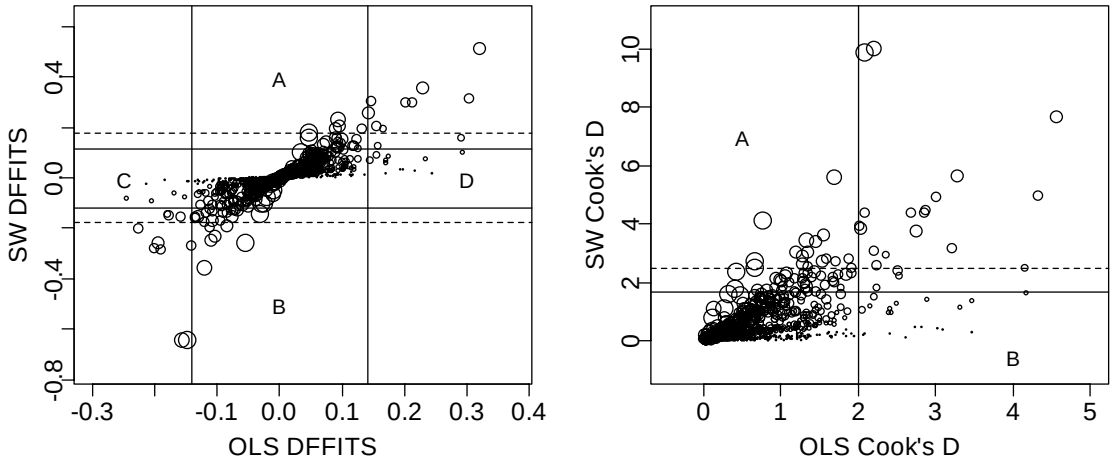


Table 5.12 numerically reports the weight discrepancies between the observations uniquely identified by either OLS or SW diagnostics. The leverage and modified Cook's Distance are more sensitive to extreme sample weights, compared to other diagnostic statistics. They tend to detect more influential points for survey data than the OLS approaches. Analysts may consider properly raising the cutoff values for these statistics in their research in order not to over-identify influential points.

Table 5.12. Number of Outliers Identified and Associated Weight Ranges for NHANES Data.

Diagnostic Statistics	Outliers Identified by OLS only		Outliers Identified by SW only	
	Counts	Weight Range	Counts	Weight Range
Leverage	24	(875.5, 13085.8)	85	(16929.6, 103831.2)
Residual	1	(2730.1, 2730.1)	8	(1791.1, 36955.3)
DFBETAS(BMI)	25	(1773.5, 23677.5)	12	(32451.1, 103831.2)
DFFITS	21	(994.9, 17366.9)	28	(29617.1, 103831.2)
Modified Cook's D	21	(994.9, 17366.9)	35	(21194.0 103831.2)

The parameter estimates after outliers were removed are listed in Table 5.13. The difference between the OLS and SW estimates and the two diagnostic schemes is trivial. The removal of observations with large DFBETAS of BMI causes the largest change in the estimated slope of BMI. The SW estimates seem to be less affected by the removal of influential points than the OLS ones. Unlike the SMHO data, the NHANES data set does not contain extremely distinct points and the outliers are spread evenly at both sides of the regression line. Hence the deletion of the identified outliers does not move the

regression line dramatically.

Table 5.13. Estimated Slopes of BMI from Full Sample and Reduced Samples by Different Diagnostic Approaches for NHANES Data.

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

*** significant at level 0.000

5.4 Simulation

A difficulty with the analysis in the previous section is that the best underlying population model is unknown. As a result, we cannot be sure whether removing influential points improves estimates or actually make them worse. Thus, it is important to study the proposed methods in a situation where the underlying model is known. To evaluate the performance of the diagnostic approaches proposed and modified in Chapter 3, we also 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. When influential points are identified, there may be several reasons and remedies. The particular situation considered in the simulation was one in which unusual, extreme values (in $\mathbf{Y}$, $\mathbf{X}$, or $\mathbf{W}$) cause observations to be influential. We generated a population in which the underlying model was known and then injected outlying observations in various ways. 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.

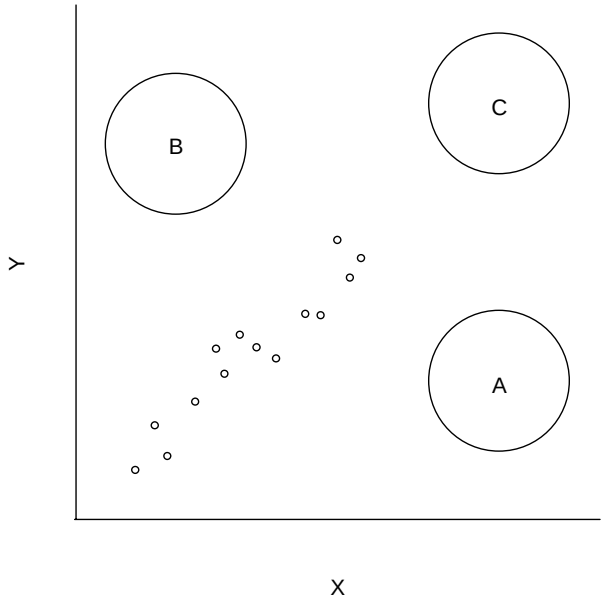
5.4.1 Description of Study Population and Sample Design

The population used in the simulation was created from the 1998 SMHO data file,

which was used for case study 1 in Section 5.2. The SMHO population has $N = 875$ observations and two auxiliary variables, number of beds and number of additions. To construct the "core" part of the study population, first we excluded observations with outlying number of beds or number of additions. The remaining 543 cases have number of beds between 10 and 300 and number of additions between 10 and 7000. A $\mathbf{Y}$ vector was then generated based on the two auxiliary variables using Gamma distributions $Y_i \sim Gamma(s, a)$, with shape parameter $s = \sigma^2 / \mathbf{x}_i^T \boldsymbol{\beta}$ and scale parameter $a = \left(\mathbf{x}_i^T \boldsymbol{\beta}\right)^2 / \sigma^2$, $\mathbf{x}_i$ is a vector including intercept, number of beds, and number of additions, $\boldsymbol{\beta} = (5000, 80, 4)^T$, and $\sigma^2 = 8 \times 10^6$. Then Y_i has a mean $\mathbf{x}_i^T \boldsymbol{\beta}$ and a constant variance σ^2.

For an OLS linear regression, an influential point may be outlying or extreme with respect to its Y value, its $\mathbf{X}$ values, or both, and it may locate either above or below the regression line. Figure 5.15 illustrates this for the case of regression with a single predictor variable. Points in areas A and B in Figure 5.15 are likely to be influential in affecting the fit of the regression function and pull the regression line to the direction where they reside. If the outliers are evenly and symmetrically scattered in the two areas, they may not change the coefficient estimates much but greatly affect the estimated standard errors. Points in area C may not be too influential if their $\mathbf{Y}$ values are consistent with the regression relation displayed by the nonextreme cases. However, they can also be influential in determining the variance estimates if the $\mathbf{Y}$ and $\mathbf{X}$ values are extremely different from other points in the data set.

Figure 5.15. Scatter Plot for Regression with One Predictor Variable Illustrating Outlying Cases.



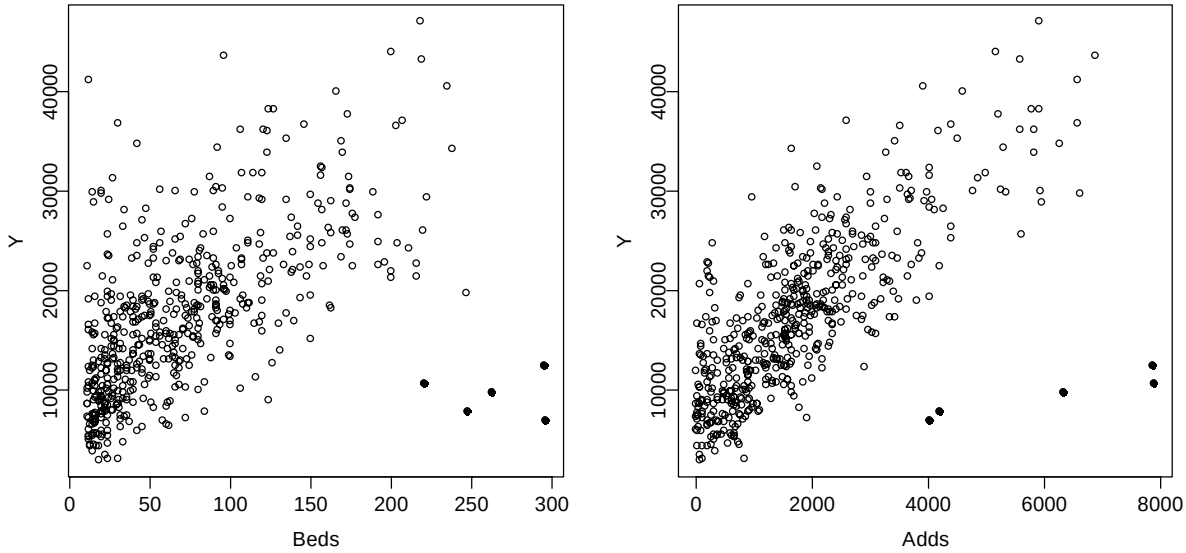
Five possible influential points, analogous to those located in area A of Figure 5.15, were created and added to the "core" population. The $\mathbf{X}$ values for the 5 outliers were generated from two uniform distributions. Number of beds was selected between 200 and 300 and number of additions was chosen between 4000 and 8000. The corresponding $\mathbf{Y}$ values were created by $\mathbf{Y} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \tilde{\sigma}^2\mathbf{I}\right)$, where $\tilde{\boldsymbol{\beta}} = \left(500,10,1\right)^T$, and $\tilde{\sigma}^2 = 10^3$. Therefore, the study population consists of three variables and has a size of 548. Figure 5.16 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. In Section 3.2 we have postulated that "the 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." According to this rule, we used the OLS estimates on the "core" population to be the "core" parameters. Table 5.14 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 5.14. Parameter Estimations Based on "Core" Population and Full Population with 5 Outliers.

Independent Variables	Underlying Core Model Parameters	Finite Population Parameters					
		Core			Full		
		Coeff.	SE	*t*	Coeff.	SE	*t*
Intercept	5000	5056.62	239.57	21.11	7099.48	363.95	19.51
# of Beds	80	76.01	2.48	30.66	56.72	3.78	15.00
# of Additions	4	4.09	0.09	43.41	3.50	0.15	24.06

Figure 5.16. Plots of Y versus Auxiliary Variables Including 5 Generated Outliers.



The samples were selected from the constructed population with probability proportional to size (PPS) 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. For each sample, there are four variables available for regression analysis: $\mathbf{Y}$, number of beds, number of additions, and sample weight.

5.4.2 Diagnostic Scheme and Regression

Since regressions will be run on both full samples and reduced samples without the identified influential cases, a scheme needs to be specified to describe which diagnostic approaches will be used and what cutoff values they will adopt. 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 leverages, residuals, DFBETAS, DFFITS, and modified 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 for DFBETAS, DFFITS, and modified Cook's Distance. When we utilized DFBETAS statistics to detect influential units, we examined 3 sets of units: (1) units with extreme DFBETAS of number of beds; (2) units with extreme DFBETAS of number of additions; and (3) units in either (1) or (2). In addition, we also grouped units which were identified by at least two diagnostic methods described above. In all, based on each selected sample, we were able to create 16 reduced samples (8 from the OLS diagnostics and 8 from the SW diagnostics). The OLS and the SW regressions were run on full samples and the corresponding regressions were run on reduced samples. We recorded 18 sets of coefficient estimates and their standard errors at each iteration of the simulation.

5.4.3 Summary Statistics

The entire sampling, diagnostic, and regression process was repeated 5,000 times in the simulation. Summary statistics across the simulation include:

1) Average number of identified outliers and average number of correctly identified outliers (Correctly identified outliers refer to those that match the 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 $relbias\left(\hat{\boldsymbol{\beta}}\right) = bias\left(\hat{\boldsymbol{\beta}}\right) \Big/ \boldsymbol{\beta} = \left(\bar{\hat{\boldsymbol{\beta}}} - \boldsymbol{\beta}\right) \Big/ \boldsymbol{\beta}$, where $\bar{\hat{\boldsymbol{\beta}}} = \sum_{i=1}^{5000} \hat{\boldsymbol{\beta}}^{(i)} \Big/ 5000$, $\hat{\boldsymbol{\beta}}^{(i)}$ is the

estimate of the parameter vector from sample i, and $\boldsymbol{\beta} = (5056.62, 76.01, 4.09)^T$ is the finite population "core" parameter vector.

3) The estimated standard errors of model parameter estimates as compared to the empirical standard errors. The average estimated standard error of $\hat{\boldsymbol{\beta}}$ was calculated as $se(\hat{\boldsymbol{\beta}}) = \sqrt{\sum_i v(\hat{\boldsymbol{\beta}}^{(i)}) / 5000}$, where $v(\hat{\boldsymbol{\beta}}^{(i)})$ is the estimated variance of $\hat{\boldsymbol{\beta}}^{(i)}$ which was calculated at the ith iteration, and $i = 1, ..., 5000$. The empirical standard error of $\hat{\boldsymbol{\beta}}$ was defined as $Se(\hat{\boldsymbol{\beta}}) = \sqrt{\sum_i (\hat{\boldsymbol{\beta}}^{(i)} - \bar{\hat{\boldsymbol{\beta}}})^2 / 5000}$.

4) The percentages of intervals that include the finite population "core" parameters at the nominal 95 percent level. The confidence intervals for $\hat{\boldsymbol{\beta}}^{(i)}$ were computed as $\hat{\boldsymbol{\beta}}^{(i)} \pm 1.96\sqrt{v(\hat{\boldsymbol{\beta}}^{(i)})}$.

These summary statistics were evaluated for each of the 18 estimate sets.

5.4.4 Simulation Results

This section presents the main results from the simulation. Table 5.15 reports the average number of units that were identified and were correctly identified as influential by each diagnostic method, either OLS or SW. By "correctly identified" we mean the influential points identified from the sample match the outliers created in the population. 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 such as residuals, DFBETAS (either), DFFITS, and modified Cook's distance. On the other hand, the SW leverages only identified less than half of the sampled population outliers since the outliers in the population were associated with very small sample weights and hence less likely to be recognized. Using residuals as the diagnostic technique, we identified fewer population non-outliers than other approaches because residual diagnostic intends to filter points that are outlying with respect to their **Y** values. The results of the SW

diagnostics showed that some points, which were not labeled 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 5.15. Number of Influential Observations Identified and Correctly Identified in Population with 5 Outliers.

Diagnostic Approaches	Average # of Outliers Identified	Average # of Outliers Correctly Identified
OLS Leverages	10.6	2.7
SW Leverages	9.1	1.4
OLS Residuals	3.5	2.9
SW Residuals	4.1	2.9
OLS DFBETAS (beds)	5.7	2.7
SW DFBETAS (beds)	4.6	2.8
OLS DFBETAS (adds)	6.6	2.3
SW DFBETAS (adds)	4.0	1.8
OLS DFBETAS (either)	9.3	2.9
SW DFBETAS (either)	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
OLS >=2 methods	7.7	2.9
SW >=2 methods	8.2	2.9
Average # of Outliers Sampled: 2.9		

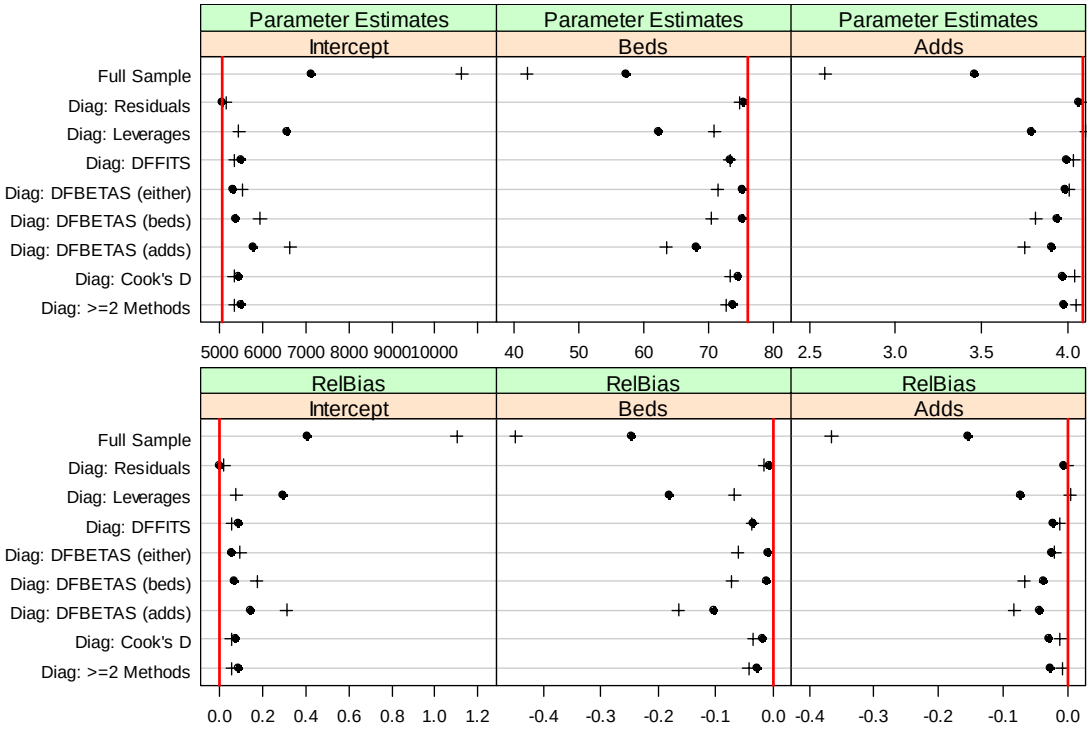
The average parameter estimates across the iterations and the relative biases, which are listed in Table 5.16 and graphed in Figure 5.17, are good indicators to gauge the effectiveness of the diagnostic methods. They also confirm the analysis we presented above. Diagnostic approaches are useful to reduce the biases in both the OLS and the SW full sample estimates with respect to the core parameters, especially when all of the population outliers were identified and deleted. The relative biases were reduced to as low as almost less than 5% for the estimated slopes. The three SW DFBETAS are more

successful in lessening the biases than the OLS DFBETAS for both slope estimates. This is likely because these statistics focus on the change in only one estimated parameter at a time and using sample weights in the construction of the statistic can accommodate for the effect of deleting a single unit on the rest of the estimated parameters. Some diagnostic techniques performed better than the others, subjected to types and positions of the outliers in the population and samples. It is expected that DFFITS and modified Cook's Distance statistics should have more stable performance regardless of outlier features because they summarize the changes in all estimated parameters and incorporate both leverages and residuals.

Table 5.16. Average Parameter Estimates and Relative Biases in Population with 5 Outliers.

	Average Parameter Estimates Over Iterations					
	Intercept	RelBias(%)	Beds	RelBias(%)	Adds	RelBias(%)
Full Sample OLS	10624.7	110.1	42.0	-44.7	2.6	-36.6
Full Sample SW	7132.2	41.0	57.3	-24.6	3.5	-15.5
OLS Leverages	5453.6	7.9	70.9	-6.7	4.1	0.3
SW Leverages	6556.3	29.7	62.3	-18.0	3.8	-7.4
OLS Residuals	5156.6	2.0	74.9	-1.5	4.1	-0.1
SW Residuals	5065.4	0.2	75.4	-0.8	4.1	-0.6
OLS DFBETAS (beds)	5943.6	17.5	70.6	-7.2	3.8	-6.8
SW DFBETAS (beds)	5393.9	6.7	75.2	-1.0	3.9	-3.8
OLS DFBETAS (adds)	6642.2	31.4	63.5	-16.4	3.7	-8.4
SW DFBETAS (adds)	5790.6	14.5	68.2	-10.3	3.9	-4.4
OLS DFBETAS (either)	5534.6	9.5	71.4	-6.1	4.0	-2.1
SW DFBETAS (either)	5330.4	5.4	75.3	-0.9	4.0	-2.6
OLS DFFITS	5355.2	5.9	73.3	-3.6	4.0	-1.4
SW DFFITS	5490.0	8.6	73.3	-3.6	4.0	-2.4
OLS Cook's D	5342.7	5.7	73.4	-3.4	4.0	-1.3
SW Cook's D	5451.7	7.8	74.6	-1.9	4.0	-2.9
OLS >=2 methods	5356.9	5.9	72.8	-4.2	4.0	-1.0
SW >=2 methods	5488.7	8.5	73.9	-2.8	4.0	-2.8

Figure 5.17. Dot Plot of Average Parameter Estimates and Relative Biases for OLS (+) Regressions and SW (•) Regressions in Population with 5 Outliers. In the upper panels the red vertical lines indicate the "core" parameter estimates. In the lower panels the red vertical lines were drawn at zero, which means unbiasedness.



Besides biases, it would also be 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 5.17. The coverage rates in Table 5.17 were calculated at a nominal 95% level. 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.

Table 5.17. Coverage Rates of 95% Confidence Intervals in Population with 5 Outliers.

	Real Coverage Rate of the 95% CI		
	Intercept(%)	**Beds(%)**	**Adds(%)**
Full Sample OLS	4	11	13
Full Sample SW	73	71	78
OLS Leverages	91	90	96
SW Leverages	86	87	95
OLS Residuals	96	96	97
SW Residuals	92	91	90
OLS DFBETAS (beds)	76	91	76
SW DFBETAS (beds)	89	91	91
OLS DFBETAS (adds)	58	56	88
SW DFBETAS (adds)	88	93	87
OLS DFBETAS (either)	90	89	95
SW DFBETAS (either)	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
OLS >=2 methods	93	92	97
SW >=2 methods	84	90	90

Table 5.16 and 5.17 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 5.18 we conclude that some of the standard errors were underestimated for the regressions on the reduced samples. The common reason of underestimating the SEs for OLS and 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. The standard variance estimates do not account for the possibility that the selected set of independent variables can differ from one sample to another, leading to underestimation (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 5.18. Empirical and Estimated Standard Errors of Parameter Estimates in Population with 5 Outliers.

	Estimated and Empirical Standard Errors								
	Intercept			Beds			Adds		
	Est.	*Emp.*	*Ratio (Est./Emp.)*	*Est.*	*Emp.*	*Ratio (Est./Emp.)*	*Est.*	*Emp.*	*Ratio (Est./Emp.)*
Full Sample OLS	1337.8	1819.9	0.74	10.7	11.6	0.92	0.4	0.6	0.71
Full Sample SW	1344.7	1086.0	1.24	11.8	9.3	1.27	0.5	0.4	1.12
OLS Leverages	789.0	942.6	0.84	6.9	7.8	0.89	0.3	0.3	1.08
SW Leverages	1281.4	1141.0	1.12	11.4	9.4	1.22	0.4	0.3	1.17
OLS Residuals	662.6	646.0	1.03	5.1	4.6	1.10	0.2	0.2	1.13
SW Residuals	783.2	820.5	0.95	6.2	6.9	0.90	0.3	0.3	0.85
OLS DFBETAS (beds)	781.8	1258.4	0.62	6.5	5.6	1.15	0.3	0.6	0.47
SW DFBETAS (beds)	775.1	897.7	0.86	5.5	6.2	0.88	0.3	0.3	0.94
OLS DFBETAS (adds)	943.0	1639.7	0.58	7.1	13.5	0.53	0.4	0.4	0.94
SW DFBETAS (adds)	920.9	938.8	0.98	9.5	8.2	1.16	0.3	0.3	0.95
OLS DFBETAS (either)	706.8	700.6	1.01	5.5	4.9	1.11	0.3	0.2	1.08
SW DFBETAS (either)	689.0	821.6	0.84	5.4	6.1	0.89	0.2	0.2	0.93
OLS DFFITS	692.8	691.2	1.00	5.3	5.1	1.05	0.2	0.2	1.13
SW DFFITS	590.9	784.7	0.75	4.8	5.7	0.84	0.2	0.2	0.88
OLS Cook's D	691.9	690.2	1.00	5.3	5.1	1.05	0.2	0.2	1.13
SW Cook's D	645.5	786.8	0.82	5.2	5.8	0.89	0.2	0.2	0.92
OLS >=2 methods	701.4	698.8	1.00	5.5	5.2	1.06	0.3	0.2	1.11
SW >=2 methods	646.6	765.5	0.84	5.2	5.7	0.91	0.2	0.2	0.94

5.4.5 Possible Masked Effect among Outliers

In Section 5.4.4 we have seen that all of the created population outliers can be fully identified when some OLS and SW diagnostic techniques were used. A natural question is what if we bring in more outliers in the constructed population. Will they mask the effects of each other and cause difficulties in influence analysis? In order to answer this question we designed another simulation in which 25 outliers were created using the same approach as we described in Section 5.4.1, and were inserted to the same core population. Figure 5.18 displays the positions of the outliers. Table 5.19 reports the estimated coefficients from the population with 25 outliers. The estimated slopes decreased even more substantially to 21.09 and 2.25 than those in Table 5.14.

Figure 5.18. Plots of Y versus Auxiliary Variables Including 25 Generated Outliers.

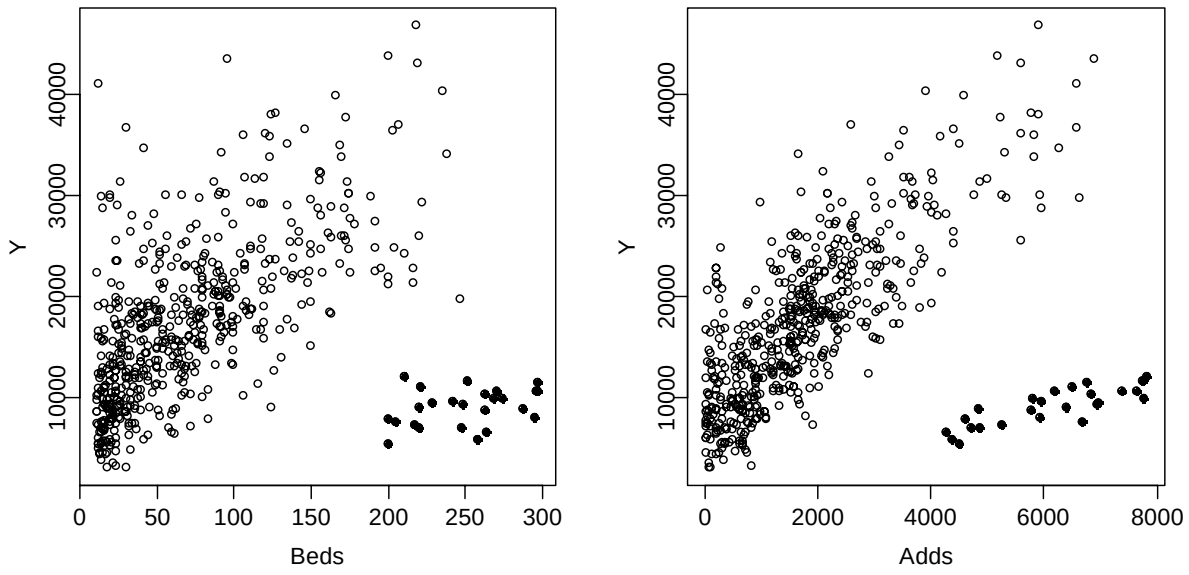


Table 5.19. Parameter Estimation Based on Population with 25 Generated Outliers.

Independent Variables	Core Model Parameters	OLS Estimation		
		Coefficient	SE	*T*
Intercept	5000	11070	486.4	22.76
# of Beds	80	21.09	5.17	4.08
# of Additions	4	2.25	0.20	11.18

The same summary statistics, as those in Section 5.4.4, were calculated for the newly-created population and presented in the following tables. As shown in Table 5.20, most of the diagnostic statistics failed to identify all outliers generated in the population. Some SW approaches, such as residuals and DFFITS, performed better than the others and detected as many as 12.4 and 10.9 population outliers out of the 12.5 outliers that were on average sampled. The SW modified Cook's Distance was greatly contaminated by the masked effects among the population outliers and can only identified very few of them. We expect that, with the unidentified population outliers used in the regressions, the estimated coefficients would be negatively biased and the confidence intervals would have lower probabilities than nominal to cover the true parameters.

Table 5.20. Number of Influential Observations Identified and Correctly Identified in Population with 25 Outliers.

Diagnostic Approaches	Average # of Outliers Identified	Average # of Outliers Correctly Identified
OLS Leverages	11.4	6.0
SW Leverages	7.4	0.6
OLS Residuals	4.5	1.1
SW Residuals	16.5	12.4
OLS DFBETAS (beds)	7.9	3.5
SW DFBETAS (beds)	7.6	4.1
OLS DFBETAS (adds)	10.7	4.7
SW DFBETAS (adds)	6.8	3.5
OLS DFBETAS (either)	15.0	7.7
SW DFBETAS (either)	12.0	6.9
OLS DFFITS	12.3	7.3
SW DFFITS	21.1	10.9
OLS Cook's D	11.8	7.1
SW Cook's D	5.1	0.9
OLS >=2 methods	13.9	7.7
SW >=2 methods	18.6	10.9
Average # of Outliers Sampled: 12.5		

Due to the incomplete identification of the population outliers, the SW diagnostics considerably reduced biases compared to OLS but did not remove them completely. For example, the relative biases of Beds estimate in Table 5.21 are -46.2% with OLS DFFITS and -23.3% with SW DFFITS; for Adds the relative biases are -54.4% for OLS DFFITS and -18.4% for SW DFFITS. The OLS reduced sample estimates are usually more biased because 1) more population outliers were not identified and hence stayed in the regression fitting; 2) population outliers have relatively small sample weights and they affect the OLS estimates more than the SW ones when used in the regressions.

Table 5.21. Average Parameter Estimates and Relative Biases in Population with 25 Outliers.

	Average Parameter Estimates Over Iterations					
	Intercept	RelBias(%)	Beds	RelBias(%)	Adds	RelBias(%)
Full Sample OLS	17004.9	236.3	6.9	-90.9	0.6	-85.6
Full Sample SW	11179.0	121.1	22.6	-70.3	2.1	-47.9
OLS Leverages	13058.3	158.2	23.1	-69.6	1.8	-54.9
SW Leverages	12733.7	151.8	25.7	-66.2	1.4	-65.5
OLS Residuals	16619.6	228.7	11.9	-84.3	0.4	-90.9
SW Residuals	5307.8	5.0	73.4	-3.4	4.0	-2.2
OLS DFBETAS (beds)	15255.6	201.7	25.8	-66.0	0.4	-90.2
SW DFBETAS (beds)	10484.9	107.3	40.7	-46.4	2.0	-51.8
OLS DFBETAS (adds)	15223.9	201.1	19.5	-74.3	0.6	-84.4
SW DFBETAS (adds)	10393.9	105.5	36.8	-51.7	2.1	-49.6
OLS DFBETAS (either)	12459.2	146.4	36.7	-51.7	1.4	-65.5
SW DFBETAS (either)	9182.5	81.6	48.6	-36.1	2.5	-37.9
OLS DFFITS	11509.2	127.6	40.9	-46.2	1.9	-54.4
SW DFFITS	7807.4	54.4	58.3	-23.3	3.3	-18.4
OLS Cook's D	11749.2	132.3	39.6	-47.9	1.8	-56.2
SW Cook's D	12404.7	145.3	25.9	-65.9	1.6	-60.4
OLS >=2 methods	11489.1	127.2	40.4	-46.8	1.9	-54.8
SW >=2 methods	7253.7	43.4	61.5	-19.1	3.4	-16.3

Applying the diagnostic methods in this population clearly does not eliminate the biases of the OLS and the SW full sample estimates. Consequently, this may have an effect on the real coverage rates of the confidence intervals. These are reported in Table 5.22. The confidence intervals based on full sample estimates almost never cover the core parameters. The coverage rates did increase after the influential observations were removed from the regressions. However, coverages with the SW diagnostics, though better than with the OLS methods, are not at a level that any analyst would consider acceptable. For example, coverage of the Beds parameter is 33% with (OLS >=2 methods) but still only 68% with (SW >=2 methods). This poor coverage is largely due to bias in the parameter estimates but also

to underestimation of standard errors as we discuss below. The SW modified Cook's Distance did not improve the coverages much because it failed to identify many population outliers.

Table 5.22. Coverage Rates of 95% Confidence Intervals in Population with 25 Outliers.

	Real Coverage Rate of the 95% CI		
	Intercept(%)	**Beds(%)**	**Adds(%)**
Full Sample OLS	0	0	0
Full Sample SW	0	1	7
OLS Leverages	2	9	14
SW Leverages	0	3	1
OLS Residuals	5	7	5
SW Residuals	91	91	88
OLS DFBETAS (beds)	1	7	1
SW DFBETAS (beds)	14	22	13
OLS DFBETAS (adds)	1	2	2
SW DFBETAS (adds)	7	9	7
OLS DFBETAS (either)	12	19	15
SW DFBETAS (either)	33	40	35
OLS DFFITS	25	35	29
SW DFFITS	47	57	65
OLS Cook's D	24	34	27
SW Cook's D	2	4	3
OLS >=2 methods	24	33	28
SW >=2 methods	62	68	71

Underestimation of the standard errors remains a problem for both OLS and SW regressions. It becomes even more severe when more outliers were created in the population and some of them were not successfully identified, as we can see in Table 5.23. In this simulation the standard errors of the OLS estimates were more underestimated than the SW ones possibly because OLS diagnostics recognized fewer population outliers than SW diagnostics.

Table 5.23. Empirical and Estimated Standard Errors of Parameter Estimates in Population with 25 Outliers.

	Estimated and Empirical Standard Errors								
	Intercept			Beds			Adds		
	Est.	*Emp.*	*Ratio (Est./Emp.)*	*Est.*	*Emp.*	*Ratio (Est./Emp.)*	*Est.*	*Emp.*	*Ratio (Est./Emp.)*
Full Sample OLS	1677.8	1382.5	1.21	14.4	12.3	1.17	0.6	0.6	1.02
Full Sample SW	1466.2	1310.8	1.12	13.8	12.8	1.08	0.6	0.6	1.02
OLS Leverages	1665.4	2548.3	0.65	16.3	18.1	0.90	0.7	0.9	0.75
SW Leverages	1570.4	1683.0	0.93	13.8	13.1	1.05	0.7	0.6	1.10
OLS Residuals	1472.4	3415.5	0.43	12.8	22.6	0.57	0.5	1.2	0.42
SW Residuals	895.4	1023.6	0.87	7.0	8.3	0.84	0.3	0.4	0.79
OLS DFBETAS (beds)	1579.8	2279.5	0.69	15.0	15.9	0.95	0.6	0.9	0.63
SW DFBETAS (beds)	1433.1	2387.5	0.60	11.5	18.5	0.62	0.6	0.9	0.66
OLS DFBETAS (adds)	1526.2	2446.0	0.62	13.4	15.7	0.86	0.6	1.0	0.61
SW DFBETAS (adds)	1450.2	1862.7	0.78	12.3	13.0	0.95	0.6	0.8	0.73
OLS DFBETAS (either)	1384.3	3880.5	0.36	12.2	20.3	0.60	0.6	1.5	0.36
SW DFBETAS (either)	1356.6	2756.6	0.49	10.6	19.7	0.54	0.5	1.0	0.52
OLS DFFITS	1328.3	4784.1	0.28	11.2	26.4	0.43	0.5	1.7	0.28
SW DFFITS	1208.8	2347.0	0.52	9.1	17.0	0.53	0.4	0.8	0.57
OLS Cook's D	1343.0	4780.1	0.28	11.4	26.5	0.43	0.5	1.7	0.29
SW Cook's D	1417.6	2046.5	0.69	12.2	15.0	0.81	0.6	0.8	0.72
OLS >=2 methods	1334.0	4586.9	0.29	11.6	25.5	0.45	0.5	1.7	0.30
SW >=2 methods	1193.6	2131.8	0.56	9.0	16.0	0.56	0.4	0.7	0.61

5.4.5 Discussion

The simulation verified the theoretical conclusion that using the SW estimator on samples without the influential cases identified by the SW diagnostic methods can obtain "better" parameter estimates than keeping those cases in the sample, where "better" means the parameter estimates are closer to the core parameters on the majority of the finite population. We anticipate that this conclusion will also hold for a multistage sampling design which may involve stratification and clustering. We are able to make a general conclusion that the use of the SW diagnostics and estimators is generally more effective than using the OLS ones. The SW diagnostics are more likely to identify the points with large sample weights. If the outliers with moderate to large weights fail to be identified, then the SW estimates can be more affected than the OLS ones and have larger biases. On the other hand, if outliers with small weights are not detected, the OLS estimates can be more biased because the outliers have more power in determining the parameter estimates.

Korn and Graubard (1995) demonstrate that the sample weights commonly affect the estimates of population means more than the estimates of association. The OLS and the SW estimation methods can have similar performance if most of the outliers can be picked up. However, the OLS estimates may be greatly different from the SW ones if (1) the sampling is done at a very different rates depending upon the outcome variable (Korn and Graubard, 1995); or (2) the model is misspecified and the omitted variable has a strong interaction with the weights (Kott 1991). Both estimators could be biased but the bias of the SW estimator decreases and may be ignored when the sample size is large. The SW regression estimator and diagnostics are recommended because they provide better protection against the model misspecification. For the SW diagnostics the change towards reducing the biases is due to two reasons: (1) the use of sample weights $\mathbf{W}$ which compensate for the unequal selection probabilities in the sample design; and (2) the removal of the units with distinct $\mathbf{Y}$, $\mathbf{X}$, or $\mathbf{W}$ values makes the regression line move closer to the "core" model.

The coefficient estimates from the reduced samples without the identified outliers may still be quite biased if there are a few outliers still not recognized. This is possibly due to the drawback of the single-case deletion methods, or more specifically, the masked effect between the outliers. The effect of deleting a possible influential point may not be correctly evaluated since other outliers, especially outliers of the same type, are still included in the sample. This problem is likely to be resolved or alleviated by using multiple-case deletion method which simultaneously removes an influential group. The simulation will be revisited using the forward search algorithm in Section 5.6.

Different SW diagnostic approaches emphasize different types of outliers and different influence measures on the regression. For example, leverages identify observations with large $\mathbf{X}$ values and weights, whereas residuals are more likely to detect cases with large $\mathbf{Y}$ values; DFBETAS statistics measure the effect of removing outliers on specific coefficients, whereas DFFITS and modified Cook's Distance are overall statistics which summarize the changes in all coefficients. It is important to scientifically and properly use one statistic or combination of statistics, choose appropriate cutoff values, and correctly define and deal with the identified outlying cases.

5.5 Case Studies Revisited: Forward Search Method

In Section 5.2 and 5.3 we have evaluated the single-case deletion diagnostic statistics by two case studies using SMHO and NHANES data sets. In this section we will revisit the two case studies and try to identify groups of influential observations by the forward search method. Individual outliers may mask each other in the sense that they will not be identified by the single-case deletion methods even though the group as a whole is influential. Through the comparison between the single-case deletion and multiple-case deletion methods, we would be able to investigate that whether there are masked outliers in the data and how the deletion of them will change the regression estimates.

5.5.1 Case Study 1 Revisited: SMHO data

(1) Selection of Initial Subsets

Initial subsets of size $m = 20$ were selected among the observations which were not identified by any of the single-case deletion methods adapted and modified in Chapter 3, such as leverages, residuals, DFBETAS, DFFITS, and modified Cook's distance. As in Section 5.2 the full sample includes 875 organizations. Out of 775 never-identified cases, we randomly picked 5000 subsamples of size 20 and kept the one with minimum median squared residuals from the SW regression as the starting subset. An additional subset of size 20 was chosen using the same approach from 5000 different subsamples. Both subsets will be used to initiate the searching in order to verify that the initial subsets are outlier-free.

(2) The Key Statistic and Other Measurements

Starting with an initial subset, the forward searching process continues by adding one new observation at a time which causes the smallest change in the parameter estimates, measured by the key statistic, in this case the delete-one version modified Cook's Distance, and ends when all units are included in the model fitting. To lessen the computational burden, extended Cook's Distance was calculated using $ED_i = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)\right)^T \left[v_L\left(\hat{\boldsymbol{\beta}}(i)\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)\right)$, where i indicated the observation newly added to the subset. During the process, MDFFIT and ED_D were also recorded since they can be helpful with drawing a line between the outliers and the non-outliers. In both statistics, the deletion set D includes the observations other than those in the subset used for model fitting.

(3) Results

The results of applying the forward search method to the SMHO data are presented numerically and graphically. Figure 5.19 shows the changes in the key statistic, single-case deletion version modified Cook's Distance, as additional observations were entered into the

subset for regression fitting. Note that the subset sizes start at 20 but the horizontal axis in the plots are truncated on the left to avoid showing the first 400 steps. Since the key statistic measures the change in all of the estimated coefficients when an extra unit is added to the regression, the inclusion of an influential observation will be signaled by a easily noticeable increase in this statistic. Although two independent forward search processes were separately conducted with different initial subsets, curves of the key statistic illustrate the same trend. After the subset size reached around 800, the modified Cook's Distance began to increase gradually. When the subset included more than about 850 observations, both curves rose dramatically. Using two multiple-case deletion statistics, Figure 5.20 and 5.21 describe alternative measurements of changes in the parameter estimates while the outliers came into the regression, the group-deletion version extended Cook's Distance ED_D and MDFFIT as defined in Section 4.1. Because these two statistics assess the difference between the estimates based on the subset and the estimates based on the full sample, we expect that they will be subjected to substantial fluctuations when some outliers begin to enter the regression and eventually have dramatic drop when all outliers come into the subset. In Figure 5.20 the group-deletion version modified Cook's Distance dropped quickly when subset size is beyond 800. Meanwhile, MDFFIT statistic tended to decrease fast but with occasional peaks in the curves. The estimated intercept and slopes are graphically displayed in Figure 5.22. Although the curves have moderate fluctuations before the subset size is near 800, they demonstrate huge increases and decreases in the estimates when the outliers began to enter the subset.

Figure 5.19. Plots of Single-Case Deletion Based Modified Cook's Distance from Forward Search with Two Different Initial Subsets in SMHO Data.

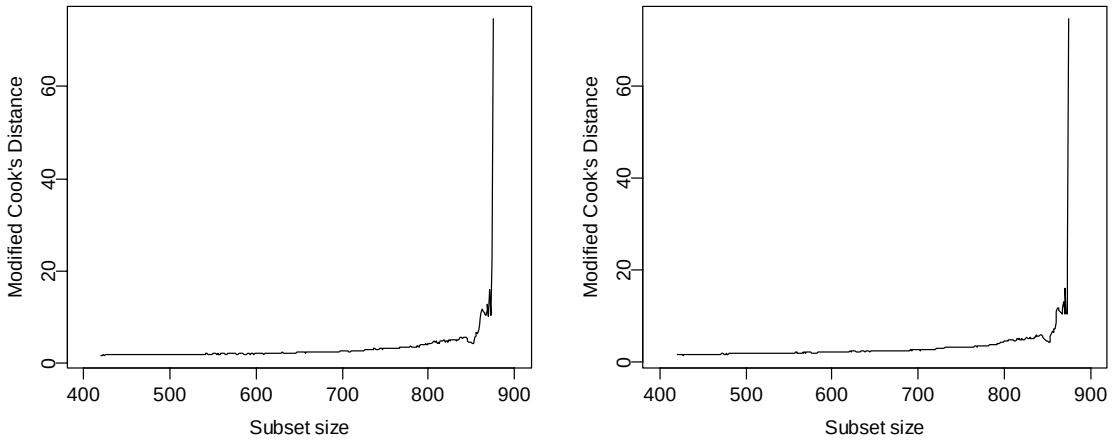


Figure 5.20. Plots of Multiple-Case Deletion Extended Cook's Distance from Forward Search with Two Different Initial Subsets in SMHO Data.

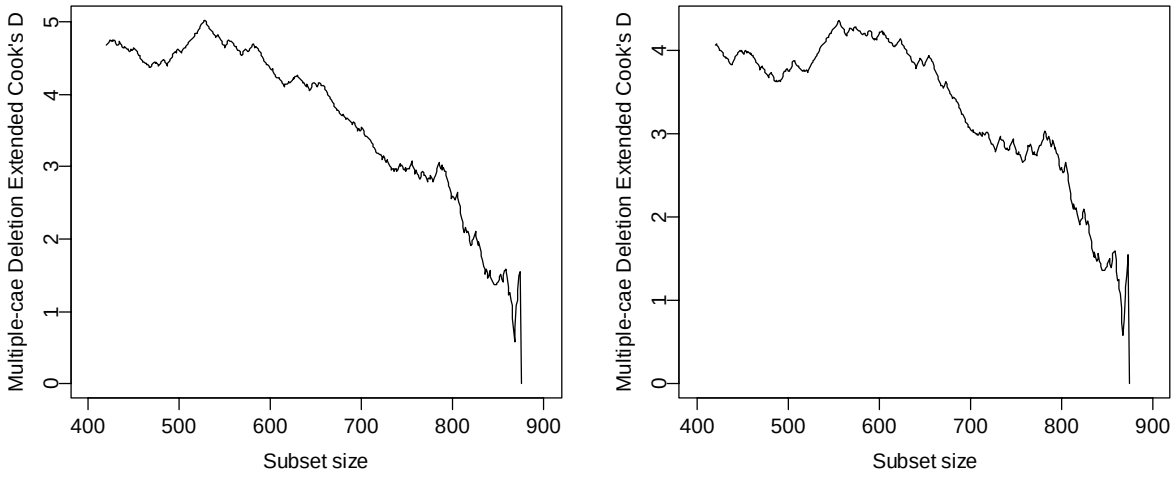


Figure 5.21. Plots of MDFFIT from Forward Search with Two Different Initial Subsets in SMHO Data.

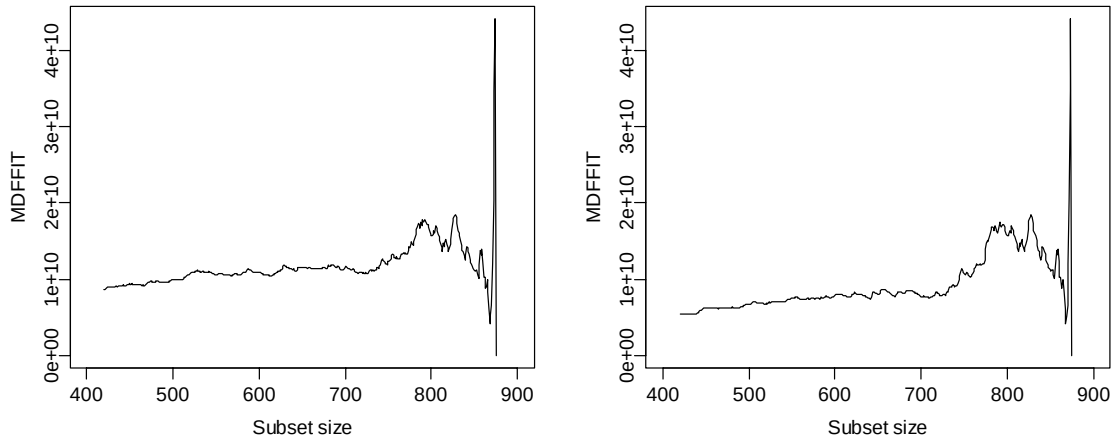
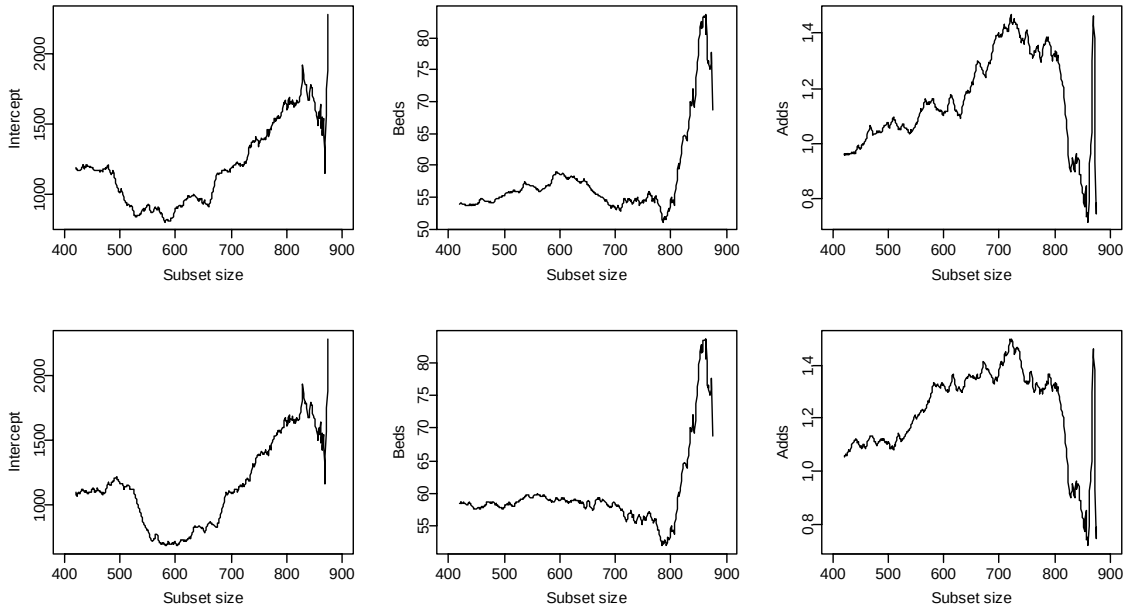


Figure 5.22. Plots of Parameter Estimates from Forward Search with Two Different Initial Subsets in SMHO Data.



All the statistics indicate that the outliers were first introduced into the subset approximately within the last 100 searching steps. The two forward searches with different starting subsets identified matching outliers after the subset size reached 792. Therefore, we determined to define an influential group containing 83 observations, among which 20 were never identified by the single-case deletion methods.

Table 5.24 reports the SW parameter estimates after the influential group was excluded. The intercept increased from 514.08 to 1612.32 and became significant. Both coefficients for number of beds and number of additions decreased radically, compared to those from the full sample. The drop in the estimated slope of number of beds is even much greater than that from samples removing outliers identified by the single-case deletion modified Cook's Distance.

The scatterplots and added variable plots in Figure 5.23 are helpful to explain the huge declines in the estimated slopes. In the plots there is an exceptionally outlying observation with extreme Y and $\mathbf{X}$ values in the upper right corner. When it is included in the sample for model fitting, even with a relatively small sample weight, it has great power in determining the regression coefficients. Moreover, it can mask the effects of the outliers above the regression line since this point itself also pulls the regression upward. Therefore, in Figure 5.5, few outliers above the regression line were identified. On the contrary, many influential points were detected in Figure 5.23, which may cause the increase in coefficient estimates when they are included in the sample.

Figure 5.23. Scatterplots with SW Scatterplot Smoothing and SW Added Variable Plots with Dark Bubbles Symbolizing Influential Points Identified by Forward Search for SMHO Data.

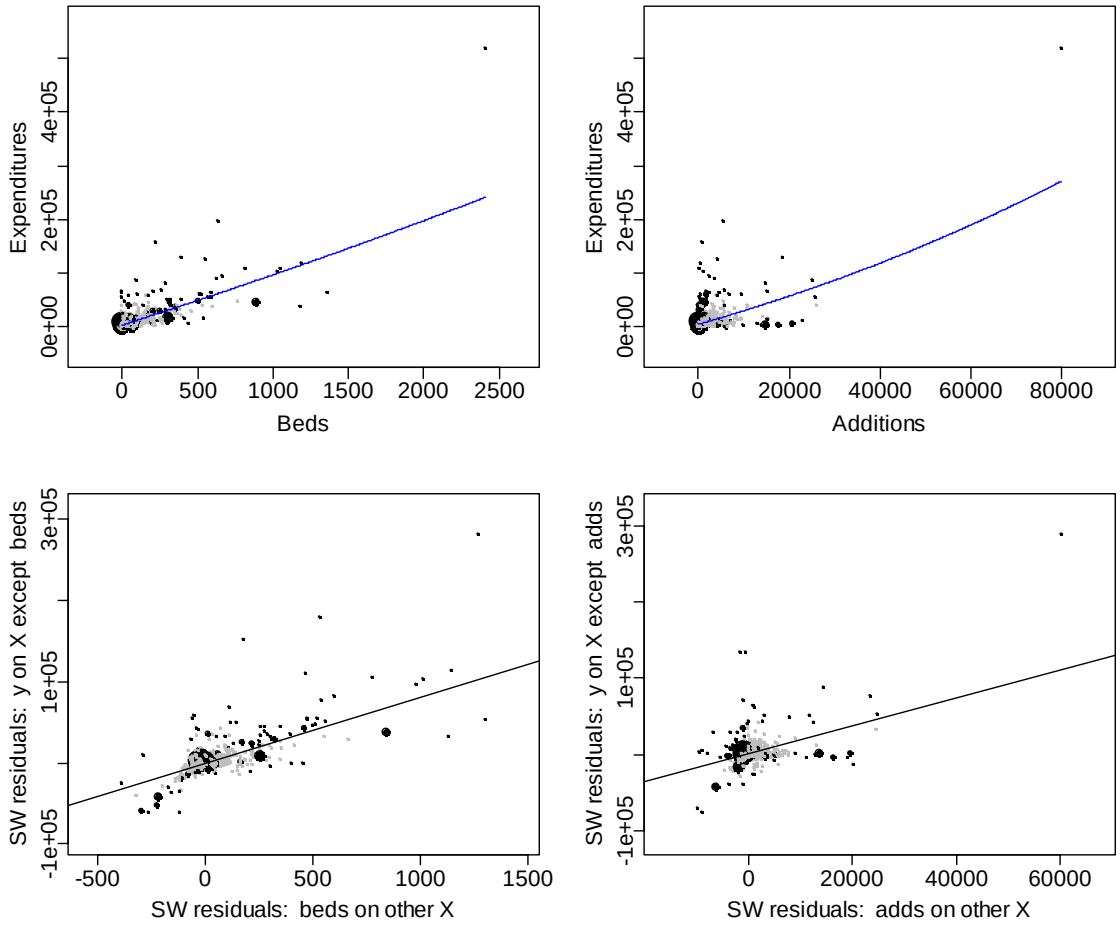


Table 5.24. Parameter Estimates of SMHO Regression after Influential Group Identified by Forward Search was Deleted.

Independent	SW full sample			SW large Cook's D			SW Forward Search		
Variables	Coeff.	SE	*t*	Coeff.	SE	*t*	Coeff.	SE	*t*
Intercept	514.08	1157.71	0.44	932.43	345.86	2.70	1612.32	181.29	8.89
Beds	81.23	13.14	6.18	82.83	5.72	14.48	52.06	2.50	20.86
Adds	1.842	0.758	2.43	1.43	0.26	5.43	1.33	0.11	12.57

(4) Discussion

The forward search method is effective to separate the outliers from the non-outliers as a

group and avoid the masked effect among outliers. It may identify a different influential set of observations and produce different parameter estimates after removing the identified influential group, compared to the single-case-deletion diagnostics. Using intuitive judgment and empirical assessment may be better than using fixed cutoffs for the key statistics when we need to filter the outliers. Different starting subsets, on one hand, are useful to verify the initial exclusion of the outliers; in addition, they are helpful in determining which points should be labeled as influential. The single-case deletion diagnostics, though having their drawbacks, form the basis of the forward search method and contribute to the choice of the initial subset and the monitor of the searching process.

5.5.2 Case Study 2 Revisited: NHANES data

The NHANES data are collected from a complex design involving stratification and clustering, which needs to be accounted for at the selection of the initial subset. Among the units which were never identified by any single-case deletion method, we drew 5000 subsamples of size 20 by randomly picking two observations from each of the 57 PSUs within the 28 strata. As before, the subsample with the smallest median of squared residuals was chosen to the initial subset. We generated two different initial subsets for the purpose of verification.

The changes in the regression were recorded by a few diagnostic statistics while new observations joined the subset. Figure 5.24 shows the variation in the key statistic, single-case deletion modified Cook's Distance. The curves increased gradually with small to moderate rises and falls before the subset size reached around 700. After that the fluctuations became stronger and rapidly increase when the subset size is larger than about 780. The points in the NHANES data are not associated with extremely distant $\mathbf{Y}$ and $\mathbf{X}$ values and are almost symmetrically and evenly distributed around the regression line. Therefore, the peaks and valleys in the Cook's Distance curves were caused by points at different sides of the regression line alternatively entering into the model fitting. Figure 5.25 and 5.26 show

abrupt decrease in group-deletion Cook's Distance and dramatic fluctuations in MDFFIT when the subset size exceeds 780 or so.

Figure 5.24. Plots of Single-Case Deletion Modified Cook's Distance from Forward Search with Two Different Initial Subsets for NHANES data.

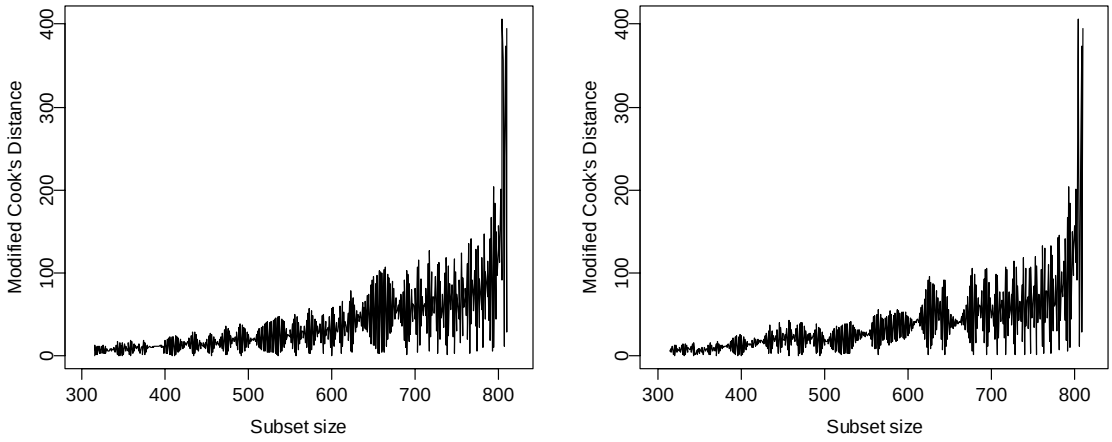


Figure 5.25. Plots of Multiple-Case Deletion extended Cook's Distance from Forward Search with Two Different Initial Subsets for NHANES data.

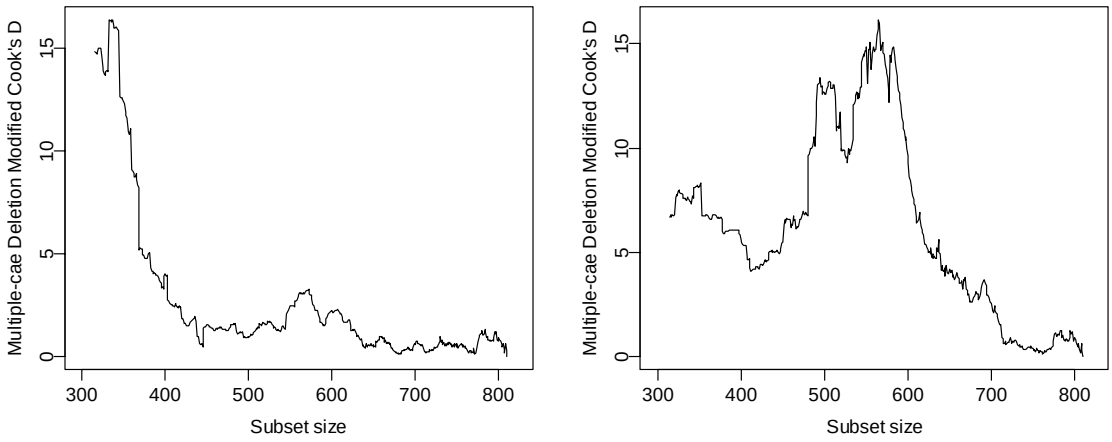
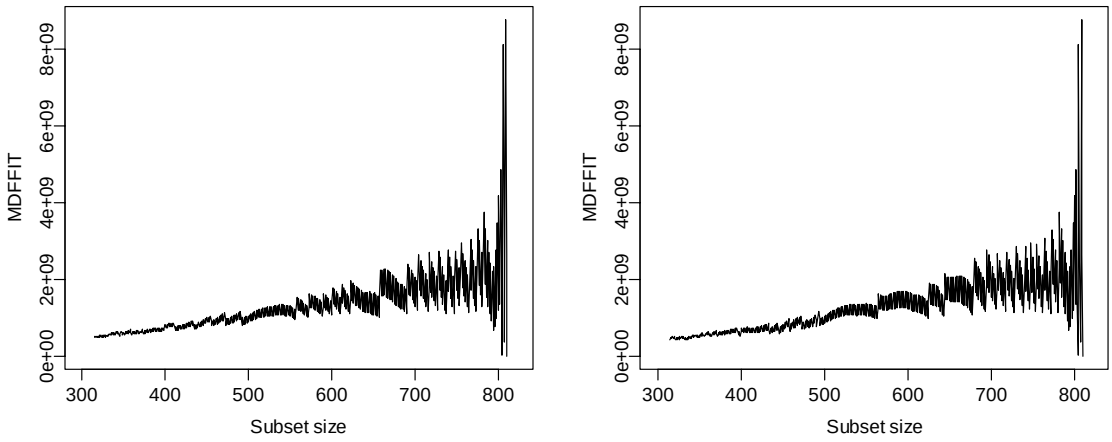


Figure 5.26. Plots of MDFFIT from Forward Search with Two Different Initial Subsets for NHANES data.



By summarizing the changes in the key statistic and other adjutant statistics, we defined an influential group of 41 points, which entered the subset at the last 41 steps during the forward searches using two different starting subsets. Figure 5.27 displays the estimated slopes of BMI while the sample size was emerging. After the outliers came into the regression, the parameter estimates tended to fluctuate around the full sample estimate, 0.45. This can be explained by the positions of the outliers relative to the regression line. In Figure 5.28, the scatterplot and the added variable plot of BMI illustrate that the effects of the outliers on the regression estimates are almost balanced out due to their distributions around the regression line. The parameter estimates after the influential group was deleted are listed in Table 5.25. They are similar to the estimation results from the full sample. The intercept and the slope of BMI remain significant. The point estimates are almost the same but the estimated standard errors become slightly smaller.

Figure 5.27. Plots of Estimated Slope of BMI from Forward Search with Two Different Initial Subsets for NHANES data.

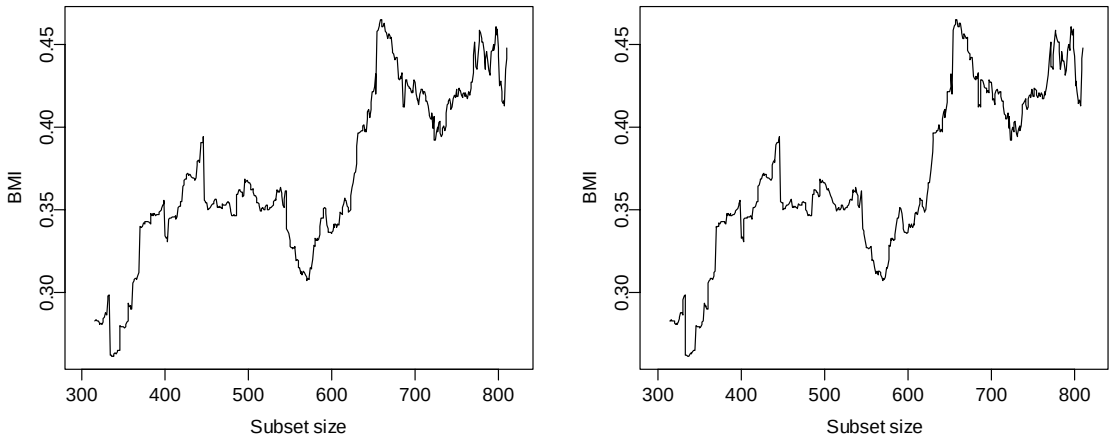
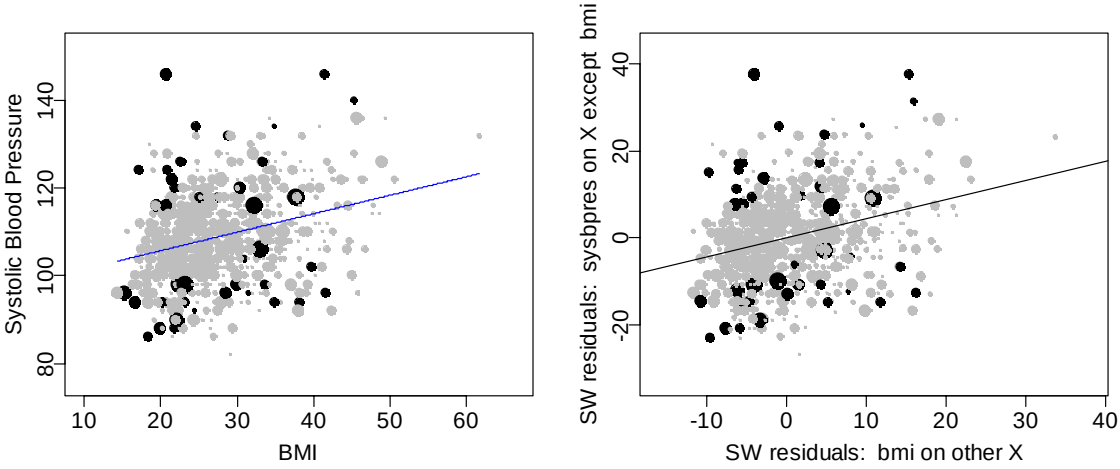


Table 5.25. Parameter Estimates of NHANES Regression after Influential Group Identified by Forward Search was Deleted.

Independent	SW full sample			SW Forward Search		
Variables	Coeff	SE	*t*	Coeff	SE	*t*
Intercept	99.79	4.72	21.16	100.68	4.22	23.89
Age	-0.15	0.17	-0.87	-0.16	0.14	-1.15
BMI	0.44	0.07	5.88	0.43	0.07	6.52
Log(Lead+1)	0.89	1.28	0.70	0.35	0.49	0.63

Figure 5.28. Scatterplot of Systolic Blood Pressure versus BMI with Scatterplot Smoothing and Added Variable Plot of BMI for NHANES data. The dark bubbles indicate points identified by the forward search method.



5.6 Simulation Revisited: Forward Search Method

The simulation in Section 5.4.5 will be revisited in this section using the forward search method because there are possibly masked effects among the outliers. We expect to obtain more correctly identified population outliers and less biased parameter estimates by applying this group-deletion method. The forward search was conducted similarly to the processes in the two case studies. However, some issues need to be reconsidered when running simulations.

(1) Selection of the initial subsets. The LMS algorithm is not convenient for filtering a "clean" starting subsample because it causes heavy computational burden. As a remedy, we applied a ranking approach. After a sample was selected, among the points that were never detected by any single-case deletion method, we assigned ranks to each observation in terms of their values of the diagnostic statistics. For example, if an observation has the smallest leverage and the second smallest residual, it is given a rank index 1 and a rank index 2. The sum of the ranking indices was calculated and 20 units with the smallest summed ranks were chosen as the starting subset. This remedial method may not enable us to acquire the best "clean" subsample of size 20, but it is likely to effectively avoid the inclusion of outliers in the selected subsample.

(2) Recognition of outliers. In the simulation we are not able to draw a line between outliers and non-outliers depending upon our case-by-case judgment and analysis. 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. This fixed cutoff may depreciate the efficiency of the forward search method to some extent.

The sampling and the diagnostics were repeated 1000 times. As in Section 5.4, we recorded the estimated parameters and their standard errors at each iteration and summarize the simulation results using statistics such as relative bias, real coverage rate of the 95% CI, empirical and estimated standard errors. These statistics were listed in Table 5.26. The biases in the parameter estimates were significantly reduced after applying the forward search

diagnostics, compared to the modified Cook's distance diagnostics in Table 5.16. The bias of the intercept dropped from 145.3% to 8.4%, and the biases of the estimated slopes decreased from -65.9% and -60.4% 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. Averaging over the iterations, it identified 18.1 influential points from the sample of size 100, and filtered 12 population outliers out of 12.5 on average sampled. Unlike those single-case deletion methods, the forward search did not falsely remove many non-outliers in the population from the regression. 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.

Table 5.26. Summary Statistics for Simulation using Forward Search Method.

	SW large Cook's D			SW Forward Search		
	Intercept	Beds	Adds	Intercept	Beds	Adds
Average Estimates	12404.7	25.9	1.6	5480.5	72.6	3.9
Relbias (%)	145.3	-65.9	-60.4	8.4	-4.5	-4.2
Coverage Rate (%)	2	4	3	75	80	78
Empirical SE	2046.5	15.0	0.8	1190.1	10.0	0.4
Estimated SE	1417.6	12.2	0.6	737.1	6.4	0.2
Est SE/Emp SE	0.69	0.81	0.72	0.62	0.64	0.65
Avg # of Outliers Sampled		12.5			12.5	
Avg # of Outliers Identified		5.1			18.1	
Avg # of Pop Outliers Identified		0.9			12.0	

Chapter 6 Conclusion

When a few or a small group of observations are different in some way from the bulk of the data, the model fitting process may be greatly affected because all observations are forced into the same regression. A premise of this research is that an analyst will be interested in estimating a model that describes the population structure reasonably well. Observations that make estimates deviate from that structure should be identified and omitted from the model fitting. In this thesis, we extended and developed a series of methods to detect and investigate observations that can be influential in determining estimates of the model parameters. Besides identifying such points or subset of points that are systematically different from the majority, we are also interested in measuring their effect on parameter estimates and inferences about models.

When using a linear regression model to analyze complex survey data, analysts usually choose the survey weighted estimator which appropriately accounts for sample weights. Hence, in survey weighted regressions, points can be influential due to combinations of outlying **Y** values, outlying **X** values, or extreme sample weights. Whether points are influential or not is affected by the fact that surveys often have fairly large sample sizes. With the incorporation of survey weights and design features, we constructed survey weighted diagnostic statistics in a way similar to the conventional OLS diagnostics. Based on the idea of case deletion, the diagnostic statistics compare the model fitting with and without possible influential points and measure the changes in the regression estimates from different aspects. Cutoff values for these statistics are determined in terms of order of magnitude analysis and distributional properties of the residuals. For survey data, we relax the traditional model assumptions such as homogeneity and independence among individual units to accommodate the sample design and features of the finite population.

Survey weighted diagnostics may identify different points than OLS diagnostics as being influential, as we have seen in the two case studies in Chapter 5. An observation with

moderate Y and $\mathbf{x}$ values may not be identified as influential by OLS approaches, but may be recognized as influential by SW methods if it is assigned an extreme sample weight. As shown numerically and graphically in Chapter 5, points identified by OLS diagnostics uniquely are usually associated with small sample weights, whereas points identified by SW diagnostics exclusively often have relatively large sample weights.

Unfortunately, techniques based on single-case deletion may not function effectively when some outliers mask the effects of others. This happens when a data set has a structure in which a group of outliers exert similar influence on the regression. Unless the group is simultaneously removed, the change in the regression can not be correctly measured because some outliers are still included in the data used to estimate the parameters. 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. It starts from a small, outlier-free subset, adds observations into the regression sequentially, and measures the fluctuations in the estimates during the search process. The group of outliers is expected to enter the model fitting at the end of the search and cause abnormal increase or decrease to the measurements of influence. The detection of outliers for this method does not completely depend on some fixed cutoff value for the statistic which monitors the change in the regression. Ideally, a decision should be made according to the trend of the statistic which is calculated at each searching step. The advantage of making case-by-case judgments is that analysts can have better control over the identification procedure. Meanwhile, it could have the drawback that the regression estimates from the reduced sample are sensitive to how many and which outliers an analyst wants to define. 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. We expected that parameter estimates from the sample excluding the identified influential units should be less biased with respect to the core population parameters. However, 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. For survey weighted diagnostics, if too many points with large sample weights are identified as influential and deleted, variance of the estimated parameters can be seriously underestimated because the variance estimators we used do not account for the variation in number of observations used in the regression. How to correct estimated standard errors to eliminate the underestimation is an open question that deserves additional research. This is a well-known problem in the model-selection literature (e.g. see Chatfield 1995) but does not appear to be addressed in research on model diagnostics.

When points are determined to be influential due to extreme survey weights, one option is to trim the weights. Potter (1990, 1993), Hulliger (1995), and Lee(1995) discussed this approach for descriptive statistics. Formal procedures for weight trimming when fitting regression models have not been explored.

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. For example, if a quadratic model is appropriate but a linear model fitted, some points may have large residuals and be identified as influential. Deleting them would be a mistake if the ability is lost to recognize that the model should be quadratic. Thus, good practice will require using more than just the diagnostics studied here.

References

Atkinson, A. C. (1982), "Regression diagnostics, transformations and constructed variables" (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological*, 44, 1-36.

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

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

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

Brewer, K. R. W., and Särndal, C. E. (1983), "Six approaches to enumerative survey sampling," in Madow, W. G. and Olkin, I. (eds.), *Incomplete Data in Sample Surveys*, vol. 3, Academic Press, 363-368.

Bruce, A. G. (1991), "Robust estimation and diagnostics for repeated sample surveys," *Mathematical Statistics Working Paper* 1991/1, Wellington, Statistics New Zealand.

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

Chambers, R. L. (1996), "Robust case weighting for multipurpose establishment surveys," *Journal of Official Statistics*, 12, 3-32.

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

Chatfield, C. (1995), "Model uncertainty, data mining, and statistical inference," *Journal of the Royal Statistical Society* A, 158, 419-466.

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

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

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

Dorfman, A. H. (1991), "Sound confidence intervals in the heteroscedastic linear model through releveraging," *Journal of the Royal Statistical Society,* Series B: Methodological, 53, 441-452.

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

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.

Ericksen, E. P. (1988), "Estimating the concentration of wealth in America," *Public Opinion Quarterly*, 52, 243-253.

Fuller, W. A. (2002), "Regression estimation for survey samples," *Survey Methodology*, 28, 5-23.

Gambino, J. (1987), "Dealing with outliers: a look at some methods used at Statistics Canada," paper prepared for the Fifth Meetings of the Advisory Committee on Statistical Methods, Ottawa: Statistics Canada.

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.

Graubard, B. I., and Korn, E. L. (2002), "Inference for superpopulation parameters using sample surveys," *Statistical Science*, 17 (1) , 73-96.

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

Hidiroglou, M. A., and Srinath, K. P. (1981), "Some estimators of a population total from simple random samples containing large units," *Journal of the American Statistical Association,* 76, 690-695.

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.

Isaki, C. T., and Fuller, W. A. (1982), "Survey design under the regression superpopulation model," *Journal of the American Statistical Association*, 77 , 89-96

Kish, L. (1995), *Survey Sampling*, New York: John Wiley.

Korn, E. L., and Graubard, B. I. (1990), "Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni statistics," *The American Statistician*, 44, 270-276.

Korn, E. L., and Graubard, B. I. (1995), "Examples of differing weighted and unweighted estimates from a sample survey," *The American Statistician*, 49, 291-295.

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

Lee, H. (1995), "Outliers in business surveys," Chapter 26 in *Business Survey Methods* (B. Cox et al., eds.) New York: John Wiley.

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

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

Miller, S. M. (1989), "Empirical processes based upon residuals from errors-in-variables regressions," *The Annals of Statistics*, 17, 282-292.

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

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

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.

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.

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

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

Pukelsheim, F. (1994), "The three sigma rule," *The American Statistician*, 48, 88-91.

Rousseeuw (1984). "Least median of squares regression," *Journal of the American Statistical Association*, 79, 871-880.

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.

Särndal, C., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley.

Shao, J. (1988), "On resampling methods for variance and bias estimation in linear models," *The Annals of Statistics*, 16 , 986-1008.

Shao, J. (1996), "Resampling methods in sample surveys," *Statistics*, 27, 203-237.

Shao, J. (1999), *Mathematical statistics*, New York: Springer-Verlag.

Srinath, K. P. (1987), "Outliers in sample surveys," paper prepared for the Fifth Meetings of the Advisory Committee on Statistical Methods, Ottawa: Statistics Canada.

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

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: John Wiley.

Weisberg, S. (1985), *Applied Linear Regression*, New York: JohnWiley.

Weisstein, E. W. (2006) "Gauss's Inequality." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/GausssInequality.html.

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.

Wolter, K. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

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