SSC Annual Meeting, May 2010
Proceedings of the Survey Methods Section

LINEAR REGRESSION DIAGNOSTICS FOR SURVEY DATA

Richard Valliant[1]

ABSTRACT

Diagnostics for linear regression models have largely been developed to handle non-survey data. The models and the sampling plans used for finite populations often entail stratification, clustering, and survey weights. In this paper we review some diagnostics that have been adapted for linear regression analysis of complex survey data. The statistics considered here include leverages, DFBETAS, DFFITS, and Cook's D. The forward search method for locating masked outliers is also illustrated. The differences in the performance of ordinary least squares and survey-weighted diagnostics are compared in an empirical study where the values of weights, response variables, and covariates vary substantially.

KEY WORDS: Complex sample, Cook's D, DFBETAS, DFFITS, Forward search, Influence, Outlier, Residual analysis.

RÉSUMÉ

Les diagnostics pour modèles de régression linéaire ont principalement été développés pour données ne provenant pas d'enquêtes. Les modèles et plans d'échantillonnage utilisés pour des populations finies impliquent souvent stratification, classification et poids de sondage. Nous adaptons certains diagnostics pour moindres carrés ordinaires ou pondérés pour données d'enquêtes. Les statistiques considérées ici incluent les leviers, DFBETAS, DFFITS et le D de Cook. Les différences de performance entre les diagnostics pour moindres carrés ordinaires et pondérés par poids de sondage sont comparées dans une étude empirique où les valeurs des poids, des variables réponses et des covariables varient substantiellement. Nous faisons aussi la revue de la méthode de recherche avant pour identifier les groupes de points influents avec des données d'enquête.

MOTS CLÉS : Analyse de résidus; D de Cook; DFBETAS; DFFITS; échantillon complexe; influence; recherche avant; valeurs aberrantes;.

1. INTRODUCTION

Diagnostics for identifying influential points are staples of standard regression texts like Belsley, et al. (1980), Cook and Weisberg (1982), Neter, Kutner, Nachtsheim, and Wasserman (1996) and Weisberg (2005). These diagnostics have been developed for linear regression models fitted with non-survey data. The diagnostic tools provided by current, popular software packages are generally based on ordinary or weighted least squares (OLS or WLS) regression and do not account for stratification, clustering, and survey weights that are features of data sets collected in complex sample surveys. The OLS/WLS diagnostics can mislead users either because survey weights are ignored, or the variances of model parameter estimates are estimated incorrectly by the standard procedures. This paper reviews some adaptations and extensions of standard regression diagnostics to survey data analysis. There has been some previous work on identifying influential points in survey data analysis. Most is geared toward outlier detection when estimating descriptive statistics like totals or means. See Li and Valliant (2009a) for a review. Work on diagnostics for models fitted from survey data is more limited. One example is Roberts, et al. (1987).

The premise in this research is that an analyst will be looking for a linear regression model that fits reasonably well for the bulk of the finite population. We have in mind two general goals. First, the influence diagnostics should allow the analyst to identify points that may not follow that model and have an influence on the size of estimated model parameters, or their estimated standard errors, or both. Second, the diagnostics should identify points that are influential

[1] Richard Valliant, Universities of Michigan and Maryland, 1218 Lefrak Hall, College Park MD 20742 USA; rvalliant@survey.umd.edu

in pseudo-maximum likelihood (PML) estimation 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.

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. Statistics include DFBETAS, DFFITS, and Cook's Distance, among others (e.g. see Neter, Kutner, Nachtsheim, and Wasserman 1996).

A key point to bear in mind is that the measures that are in the literature for non-survey regressions and the ones we present mainly have heuristic justifications only. There is limited distribution theory to support the setting of cutoff values for statistics that gauge whether a point is influential or not. Nonetheless, the measures in this paper give some practical, exploratory tools for identifying points for further examination. Section 2 introduces notation for linear regression with survey data. Section 3 reviews several diagnostics for use with complex survey data. The fourth section describes a version of the forward search method for survey data. Interspersed in sections 3 and 4 are some numerical illustrations. Section 5 is a conclusion.

2. LINEAR REGRESSION ESTIMATION WITH COMPLEX SURVEY DATA

One method of estimating parameters in linear regression using complex survey data is the pseudo maximum likelihood (PML) approach. The first step of this approach is to write down and maximize the likelihood when all finite population units are observed. Suppose that the underlying structural model is the fixed-effects linear model given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,,\ \boldsymbol{\varepsilon} \sim N\left(0, \sigma^2\mathbf{V}\right), \qquad (1)$$

with $\mathbf{Y} = \left(Y_1,\ldots,Y_n\right)^T$, $\mathbf{X} = \left(\mathbf{x}_1,\ldots,\mathbf{x}_n\right)^T$, $\mathbf{x}_i^T = \left(x_{i1}, x_{i2},\ldots,x_{ip}\right)$, $\boldsymbol{\varepsilon} = \left(\varepsilon_1,\ldots,\varepsilon_n\right)^T$, and $\mathbf{V} = diag\left(v_i\right)$ is an $n \times n$ diagonal matrix.. The pseudo maximum likelihood estimator (PMLE) of $\boldsymbol{\beta}$, assuming normal errors, is $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{Y}$. In this paper, a model with cluster-correlated errors is not considered, although extensions to that case are in Li (2007). If we assume $\mathbf{V} = \mathbf{I}$, the PMLE reduces to $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$. This estimator will be referred as a survey weighted (SW) estimator in the following discussion and is the one usually computed by software packages that handle survey data.

3. ADAPTATIONS OF STANDARD TECHNIQUES TO SURVEY REGRESSIONS

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

3.1 Variance Estimators

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

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

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

$$\hat{\sigma}^2=\sum\nolimits_{i\in s} w_i e_i^2\Big/\left(\hat{N}-p\right) \qquad (3)$$

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

The model-based estimator above is useful because it explicitly shows the estimates of model parameters. With some simplifications, described in Li and Valliant (2009b), (2) is helpful in setting heuristic cutoffs that can be used with the diagnostics.

3.2 Leverages

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

As an example, consider the 1998 Survey of Mental Health Organizations (SMHO) used in Li and Valliant (2009a). This sample consisted of 875 facilities which were selected in a single stage with probabilities proportional to a measure of size defined as number of episodes (i.e., number. of patients at beginning of year plus number of additions (Adds) during the year.

Table 1. Quantiles of Variables in SMHO Regression

	Quantiles				
Variables	Minimum	25%	50%	75%	Maximum
Expenditure (1000’s)	16.6	2,932.5	6,240.5	11,842.6	519,863.3
# of Beds	0	6.5	36	93	2,405
# of Additions	0	558.5	1,410	2,406	79,808
Weights	1	1.42	2.48	7.76	158.86

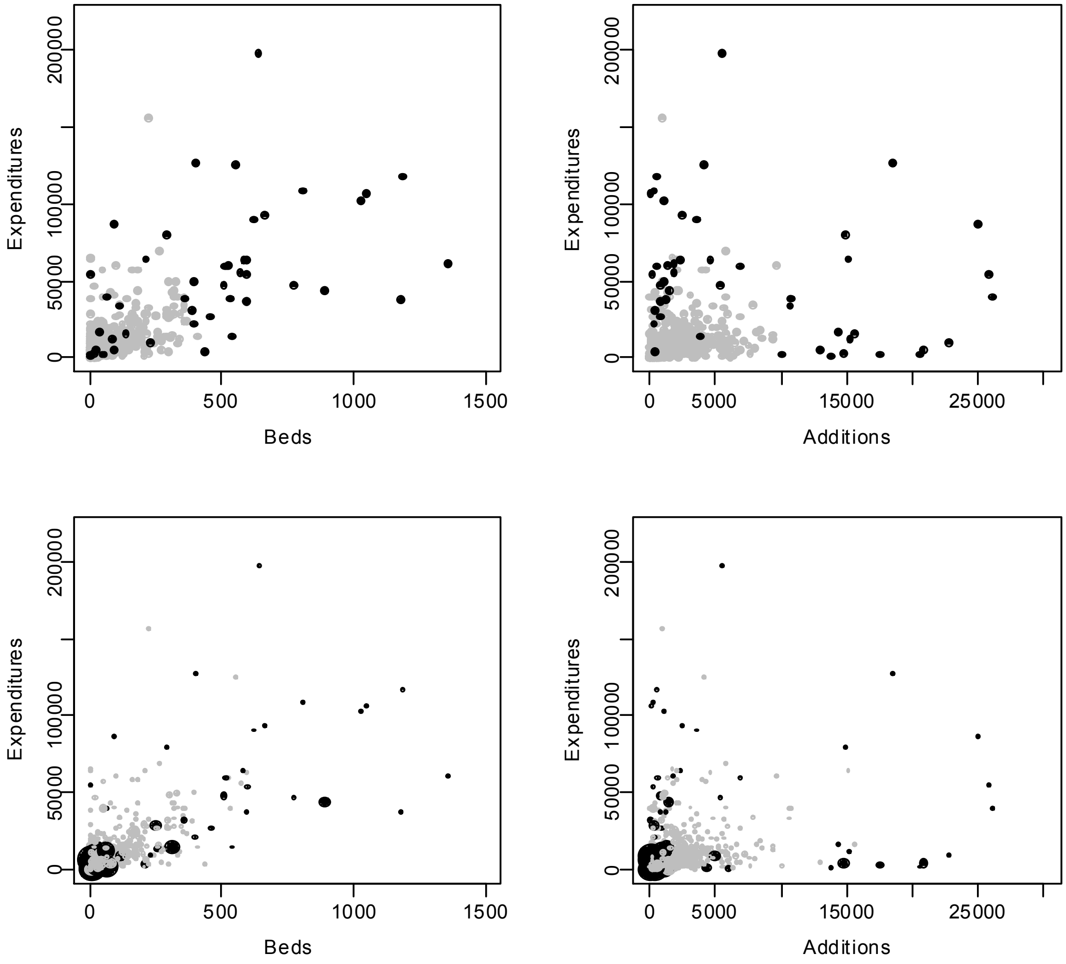
The model is to regress Expenditures on Beds and Adds. Table 1 shows some quantiles and the minima and maxima for these variables and for the survey weights. Figure 1 shows scatterplots of Expenditures vs. the two predictors. High leverage points based on OLS are highlighted in the top row; ones with high leverage based on SW are in the bottom row. One rule-of-thumb is that leverages greater than twice the average should be examined, which in this example means $h_{ii}>2p/n=0.007$. In Figure 1, 48 points have leverages based on OLS while 61 do based on SW. In addition, the points that are identified are different in OLS and SW. Note that many points near the origin are high leverage in SW because they have large weights—not outlying *Y*’s or *x*’s.

3.3 Residual Analysis

Standardizing residuals is helpful so that their variance is approximately 1. In the OLS case, a residual is scaled either by $\sqrt{\text{MSE}}$ or by its estimated standard error. Under model (1), the residual for unit i based on the PMLE is $e_i=Y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$

and its model variance is $E_M\left(e_i^2\right)=\sigma^2\left[\left(1-h_{ii}\right)^2+\sum_{i'\neq i}h_{ii'}^2\right]$. Since $h_{ii'}=O\left(n^{-1}\right)$, the term in the brackets has the form $1+o(1)$, and $E_M\left(e_i^2\right)\square\,\sigma^2$. We can standardize the residual for unit *i* as $e_i/\hat{\sigma}$ and compare it to percentiles from the distribution of a standard normal random variable. If e_i is not normal, the Gauss inequality (Pukelsheim 1994) is useful for setting a cutoff value. This inequality states that 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 $\tau^2\equiv\sigma^2+\left(\mu-\mu_0\right)^2$.

Figure 1. Scatterplots of expenditures versus beds and additions. High leverage points based on OLS (SW) are highlighted in top (bottom) row. The second row has bubbleplots with areas proportional to the survey weight.



Assume that a residual has a symmetric distribution with its mode and mean at zero. The Gauss Inequality implies that the absolute value of a residual has about 90% probability to be less than twice its standard deviation and about 95% probability to be less than three times its standard deviation. If we rescale the residuals by a consistent estimate of σ, we can use either 2 as a loose cutoff or 3 as a strict one to identify outlying residuals, depending on an analyst's preference.

Appealing to a model is necessary when analyzing residuals because it is not feasible to define the distribution of residuals from the design-based point of view, even asymptotically. For example, in single-stage sampling, $e_i=Y_i\left(1-h_{ii}\right)+\sum_{i'\neq i\in s}h_{ii'}Y_{i'}$. 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, if we approach the analysis with a working model in mind, plots of residuals are helpful in highlighting data points suspected of unduly affecting the fit of regression. For instance, plots of observed Y 's or residuals against predicted values are still useful.

3.4 DFBETAS

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

Using $DFBETA_{ij} = \left(\mathbf{A}^{-1}\mathbf{x}_i e_i w_i\right)_j \big/ (1-h_{ii}) = c_{ji} e_i \big/ (1-h_{ii})$ where $\mathbf{C} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W} = \left(c_{ji}\right)_{p\times n}$ and a variance estimator, $v\left(\hat{\beta}_j\right)$, for $\hat{\beta}_j$, a scaled statistic DFBETAS can be constructed as in the OLS case. Li and Valliant (2009b) propose a specification of DFBETAS statistic as

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

An observation i may be identified as influential on the estimation of $\hat{\beta}_j$ if $\left|DFBETAS_{ij}\right| \geq z\big/\sqrt{n}$ for z = 2 or 3. An *ad hoc* alternative would be to use a cutoff of $t_{0.025}(n-p)\big/\sqrt{n}$ where $t_{0.025}(n-p)$ is the 97.5 percentile of the *t*-distribution with $n-p$ degrees of freedom.

3.5 DFFITS

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

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

where $v\left(\hat{\beta}_j\right)$ is appropriate to the design and/or model. The model variance is again convenient for motivating cutoffs. Based on arguments in Li and Valliant (2009b), the cutoff value can be set to $z\sqrt{p/n}$ (z = 2 or 3) for using DFFITS to determine the influential observations.

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

A measure of distance from $\hat{\boldsymbol{\beta}}(i)$ to $\hat{\boldsymbol{\beta}}$ for survey data can be constructed similar to a Wald Statistic, depending on the regression model of interest and the sampling design for the survey data. An extended version of Cook's D (Cook 1977) for survey estimates is

$$ED_i = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)\right)^T \left[v\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)\right). \tag{4}$$

If $\hat{\boldsymbol{\beta}}(i)$ were replaced by an arbitrary value $\mathbf{b}_0$, (4) would have the form of a confidence ellipsoid. The new statistic ED_i can be compared to a Chi-square distribution. If ED_i were exactly equal to the $100(1-\alpha)\%$ quantile of the Chi-square distribution with *p* degrees of freedom, then the deletion of the i^{th} case would move the estimate of $\boldsymbol{\beta}$ to the edge of a

$100(1-\alpha)\%$ confidence ellipsoid based on the complete data. A large value of this quadratic form indicates that the i^{th} observation is likely to be influential in determining joint inferences about all the parameters in the regression model. Another formulation of the extended Cook's Distance can be derived from the Wald *F* statistic (Korn and Graubard 1990) as $ED_i' = \frac{n-p+1}{np}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)\right)^T\left[v\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)\right)$ and its value can be compared with quantiles from an *F* distribution.

Like the Cook's Distance, the 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 tail percentiles of *F* and Chi-square statistics are adopted as cutoffs. Following Atkinson (1982), we modify the extended Cook's Distance to be $MD_i = \sqrt{nED_i/p}$. This modified Cook's D, can be judged in terms of a standard normal distribution, implying that we can use 2 or 3 as the cutoff value.

Table 2. Effects on coefficients, standard errors, and *t*-statistics of different criteria for removing points in the SMHO regression

Independent	OLS Estimation			SW Estimation		
Variables	Coef	SE	*t*	Coef	SE	*t*
Intercept	-1201.7	526.2	-2.3	514.1	1157.7	0.4
# of Beds	94.2	3.0	31.1	81.2	13.1	6.2
# of Additions	2.3	0.1	18.5	1.8	0.8	2.4
(i) Deleting units with leverages greater than 2*p*/n=0.007						
No. deleted	48			61		
Intercept	2987.6	490.5	6.1	1993.9	353.7	5.6
# of Beds	69.3	4.3	15.9	75.8	6.8	11.2
# of Additions	0.9	0.2	4.7	1.0	0.2	4.7
(ii) Deleting units with abs standardized residuals > 3						
No. deleted	17			37		
Intercept	645.8	311.6	2.1	1674.7	386.3	4.3
# of Beds	84.5	2.0	42.7	76.2	5.3	14.4
# of Additions	1.5	0.1	14.9	0.9	0.2	4.3
(iii) Deleting units with modified Cook's D > 3						
No. deleted	44			10		
Intercept	1660.5	335.5	4.95	932.4	345.9	2.7
# of Beds	80.9	2.4	33.2	82.8	5.7	14.5
# of Additions	1.2	0.1	9.7	1.4	0.3	5.4

For illustration, we again use the 875 facility SMHO sample described earlier and regress Expenditures on Beds and Adds. Table 2 shows the changes in coefficients, standard errors (SEs), and *t*-statistics that occur after deleting units in the SMHO sample based on large leverages, standardized residuals, and modified Cook's D. Note that estimated coefficients do differ depending on the criterion used for deletion. Standard errors do decrease when outlying points are dropped as might be expected. Another point to note is that the SW deletion results differ from those for OLS, illustrating that specialized diagnostics are needed for survey-weighted least squares.

4. FORWARD SEARCH

Some outliers may be "masked" by others, i.e., they are not identifiable based on full sample diagnostics but if the points are removed, parameter estimates will change noticeably. Atkinson & Riani (2000) developed a method called *forward search* that is designed to locate masked outliers and to identify groups of outliers. A version of forward that is modified for survey data is:

(1) Choose initial subsample of size m, free of outliers. Set m = 10 or 15 times the number of parameters in model. One method of finding an outlier-free subsample is the following. Select many subsamples that reflect full sample design. Fit least median of squares (LMS) regression to each. Retain the sample with smallest median of squared WLS residuals.
(2) Compute the survey-weighted $\hat{\boldsymbol{\beta}}$ from the subset of m and the residuals for all units in full sample.
(3) Select the m+1 cases from the full sample with the smallest values of $\sqrt{w_i}e_i$. (Since SWLS minimizes $\sum\left(\sqrt{w_i}e_i\right)^2$, these are the cases that contribute the smallest amount to that criterion.)
(4) Compute SWLS $\hat{\boldsymbol{\beta}}$ from the m+1 cases and the residuals for full sample.
(5) Select the m+2 cases with smallest $\sqrt{w_i}e_i$. (Some cases from step m+1 may drop out.)

Steps (3)-(5) are repeated until the entire full sample is included. As the search unfolds, various statistics are tracked: $\hat{\boldsymbol{\beta}}$'s, $\hat{\sigma}^2$, scaled residuals, t-stats, R^2, modified Cook's D, etc. This creates series of these different evaluation diagnostics. We then look for breaks in the series which signal that outliers may have entered into the subset.

As an example, we use the 665 facility subset of the SMHO data that have non-zero inpatient beds. The regression model predicts annual expenditures based on beds, patients added during year, and an indicator for organization type (general hospital, multi-service or substance abuse, psychiatric, residential care, and veterans). Figure 2 plots $\hat{\sigma}^2$, R^2, the maximum absolute studentized residual, and the minimum absolute deletion residual at each step of search. A deletion residual for a unit excluded from the current subsample is defined as $r_{ti}^* = e_{ti}\Big/\sqrt{\hat{\sigma}^2 + \mathbf{x}_i^T v\left(\hat{\boldsymbol{\beta}}_t\right)\mathbf{x}_i}$; e_{ti} is the residual for excluded unit i at step t and the denominator is an estimate of the prediction variance of $y_{ti} - \hat{y}_{ti}$. One criterion for judging whether a standardized residual is large is to set a Bonferroni-adjusted cutoff value. In this case 3.43 is appropriate for 665 comparisons to give overall error rate of 0.2. The third and fourth panels in Figure 2 have horizontal reference lines drawn at 3.43.

Toward the end of the search, fairly sharp changes occur in each of the plots. The value of R^2 is 0.54 at the next-to-last step; it jumps to 0.67 at the last step when the Multi-service/substance-abuse unit with the largest expenditures of $520 million and 2,405 beds is added. Between steps 636 and 637, 657 and 658, 662 and 663, there are fairly large increases in the maximum studentized residual, which is a signal that an outlier has entered the subset. These jumps correspond to the entry of the following cases:

Expenditures	Beds	Adds	Type	Weight
57,682,498	157	4,021	Psychiatric	1
125,114,267	552	4,124	Psychiatric	1.454
156,376,651	222	969	Psychiatric	1

These expenditures are well above the majority for Psychiatric hospitals. The units do, however, have small weights. Figure 3 plots the tracks of standardized residuals at each step of the search. There are a number of points that are outliers throughout most of the search, but become non-outliers after step 620—these are outliers that are masked in the full sample. Conversely, there are units (marked in red) that are not outliers through most of the search but appear to be outliers in the full sample. Figure 4 plots the full sample residuals vs. a sequence number; hospitals of different type are marked by color. The masked outliers are shown in larger circles in cyan. There are a number of points that would be identified as extreme based on standardized, full sample residuals, but the masked outliers would escape attention.

[redacted]

Figure 2. Forward plot of $\hat{\sigma}^2$, R^2, max abs studentized residual, and min abs deletion residual at each step of search. Horizontal reference lines at 3.43 in the second row panels.

5. CONCLUSION

Although diagnostics are included in most software routines that compute weighted least squares estimates, they are not always appropriate for survey data analyses. Existing regression diagnostic procedures need to be adapted for use with survey data. Among the sources of influence on survey-weighted linear regressions are outliers among the Y's, extreme x's (leverage points), weights, and combinations of all of these. Although single points may be unlikely to be influential in survey data sets, groups of points may be. Such groups can be identified with the forward search method.

Two of the key questions that an analyst should consider are: What is an outlier vs. a point that just indicates high variance? What do we do once points are identified? One option is to use automatic deletion procedures, e.g., if a standardized residual is extremely large, delete the point and refit the regression. However, such automatic procedures can create inferential problems akin to those that are well known for stepwise regression. In particular, estimated standard errors based on the usual full sample formulas do not account for uncertainty in deletion methods. This is one of several areas in the analysis of survey data that need further research.

ACKNOWLEDGEMENT

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

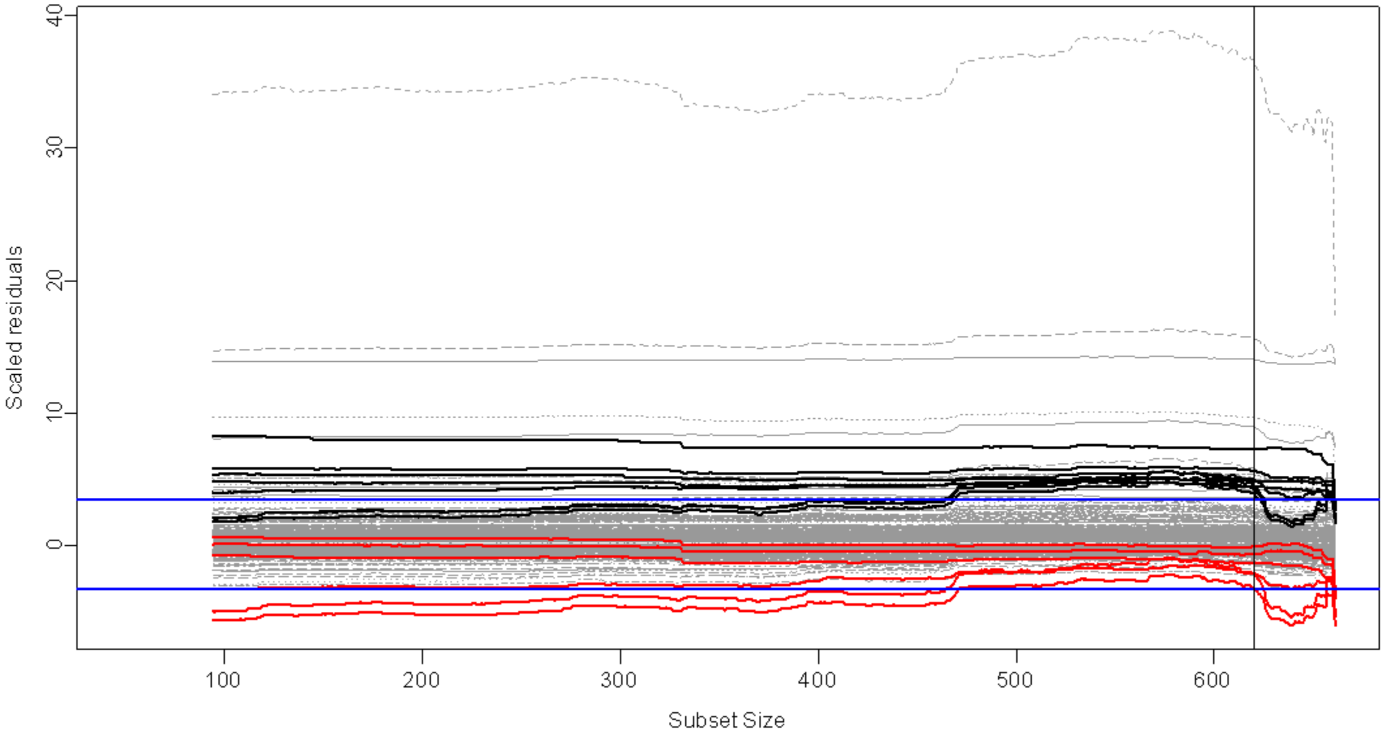


Figure 3. Residual tracks at each step of search. Horizontal reference lines at ±3.43. Vertical reference line at step 620. Heavy black lines are outliers ($|e_i/\hat{\sigma}| \geq 3.43$) at step 620 but not in full sample, step 665. Red lines are outliers in full sample but not at step 620.

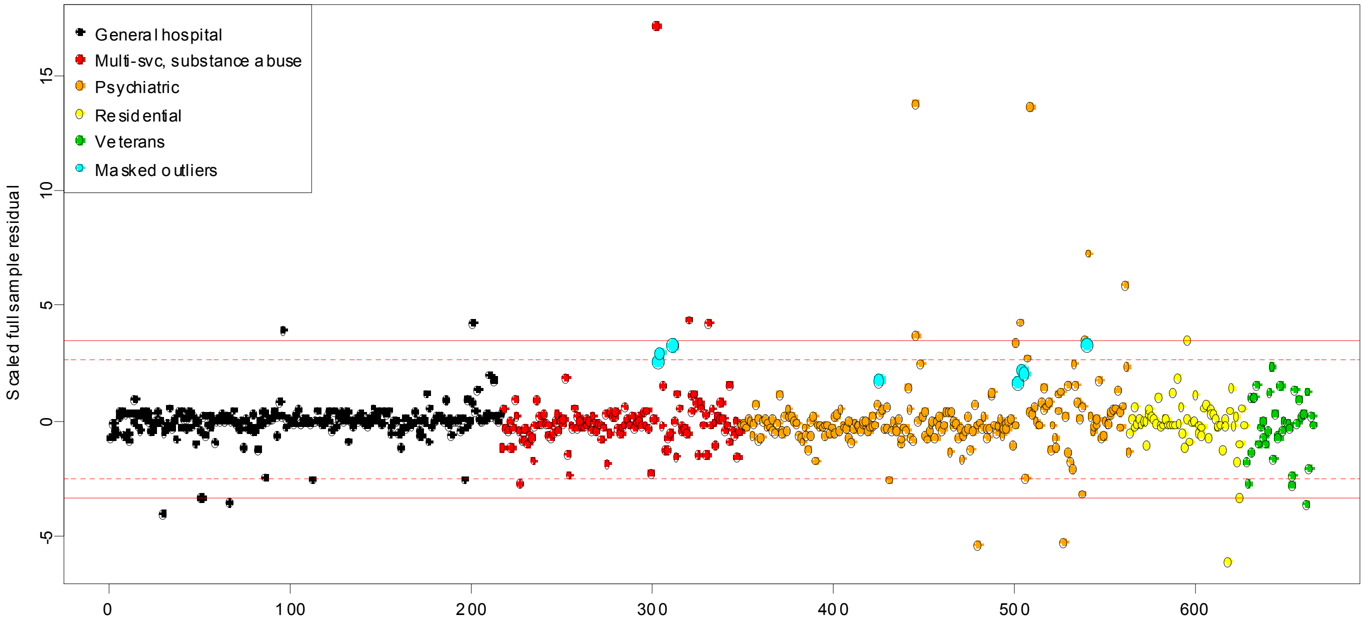


Figure 4. Full sample residuals sorted by organization type. Horizontal reference lines are drawn at ±3.43 and ±2.58 (0.005 & 0.995 normal percentiles).

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.

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.

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.

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

Li, J. and Valliant, R. (2009a). Survey Weighted Hat Matrix and Leverages. *Survey Methodology*, 15-24.

Li, J., and Valliant, R. (2009b). Linear Regression Diagnostics for Unclustered Survey Data, submitted for publication.

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

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

Roberts, G., Rao, J.N.K., and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, **74**, 1-12.

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