Sociedad de Estadística e Investigación Operativa
Test (2005) *Vol.* **14**, *No.* 2, pp. 515–542

Influence Diagnostics in Regression with Complex Designs Through Conditional Bias

M. Dolores Jiménez–Gamero
Dpto. Estadística e Investigación Operativa
Universidad de Sevilla, Spain

Juan Luis Moreno–Rebollo*
Dpto. Estadística e Investigación Operativa, Universidad de Sevilla, Spain
and Centro Andaluz de Prospectiva, Junta de Andalucía, Spain

Juan M. Muñoz–Pichardo and Ana M. Muñoz–Reyes
Dpto. Estadística e Investigación Operativa
Universidad de Sevilla, Spain

Abstract

One of the areas of Statistics in which the influence analysis has been widely studied is the multiple linear regression model. Nevertheless, the influence diagnostics proposed in this context cannot be applied to regression in complex survey, under randomized inference, since the i.i.d. case does not incorporate any probability weighting or population structure, such as clustering, stratification or measures of size into the analysis.

In this paper we introduce some influence diagnostics in regression in complex survey. They are built on the conditional bias concept (Moreno-Rebollo et al., 1999). We emphasize the similarities and differences of the proposed measures with respect to the existing ones for the i.i.d. case.

Key Words: Conditional bias, influential observation, design-based regression, survey sampling.

AMS subject classification: 62D05, 62J20.

1 Introduction

An important aspect in any statistical analysis is the study of the sensitivity of its conclusions to perturbations of the assumed model. In general terms,

*Correspondence to: Juan L. Moreno–Rebollo. Dpto. de Estadística e I.O., Facultad de Matemáticas, C/ Tarfia s.n., 41012 Sevilla, Spain. E-mail: jlmoreno@us.es

Research partly supported by the Spanish MCYT grant BFM2001-3844 and Spanish MEC grant MTM2004-0111433

Received: June 2003; Accepted: April 2004

this is the objective of the influence analysis. A large number of papers on influence analysis are centered on the study of diagnostics that are intended to detect observations impacting notably on the conclusions, in the sense that its presence or absence can cause a considerable effect on the inference. Such observations are known as influential observations.

Most influence diagnostics have been developed to the linear regression model, under the usual hypothesis, that from now on, we will call model-based regression. Some classic books on this topic are Belsley et al. (1980), Cook and Weisberg (1982) and Chatterjee and Hadi (1988). One could think of applying the diagnostics proposed in this context to study the influence in regression in complex surveys under design-based inference, that from now on we will call design-based regression. Nevertheless, when the sample is obtained from a complex survey the assumed hypotheses in model-based regression are not satisfied. In addition, model-based regression does not incorporate any probability weighting or population structure, such as clustering, stratification or measures of size, into the analysis. Therefore, as Smith (1987) affirms "conventional model-based influence diagnostics do not have immediate application to randomization inference for sample survey".

In the context of sampling from finite populations under design-based inference, the influence analysis has been scarcely treated. One of the first works in this area is that of Smith (1987). Other authors have obtained influence diagnostics in survey sampling by applying to this field ideas previously employed for the i.i.d. case. In this line we can cite the works by Gwet and Rivest (1992), Hulliger (1995), Deville (1999) and Moreno-Rebollo et al. (1999). None of these papers deal with influence in design-based regression.

The main purpose of this paper is to propose influence diagnostics in design-based regression.

With this aim we have organized the paper as follows. In Section 2 we consider design-based regression, emphasizing the differences with respect to model-based regression. Since most of the influence measures in model-based regression are functions of a measure of leverage and some type of residual (see for example Barrett and Ling, 1992; Caroni, 1987; Chatterjee and Hadi, 1986), in Section 3 we introduce analogues of these measures for design-based regression. In next section, we will see how the proposed diagnostics combine them.

By using the concept of conditional bias, Muñoz Pichardo et al. (1995) and Muñoz Pichardo et al. (2000) have obtained, in a unified way, a large number of model-based influence measures for the general linear model, which were previously proposed by several authors, each using a different argument. By properly adapting the approach followed by these authors we introduce influence measures for design-based regression. With this aim, in Section 4 we first calculate the conditional bias for our problem and compare the obtained results with its counterpart in model-based regression. Second, since the conditional bias is a population parameter, to obtain influence measures from it, we need an estimator. We consider two estimators. Third, as the conditional bias and therefore its estimations are q-vectors, we propose several influence diagnostics by normalizing the estimators previously considered, so that the observations can be ordered in a meaningful way.

In Section 5, to illustrate the proposed diagnostics we apply them to two examples: an artificial data set and a real data set. Finally, the last section summarizes and highlights the contributions of the paper.

Before ending this section we introduce some notation. Let $U = \{u_1, ..., u_N\}$ be a finite population and let $\mathbf{m} = \{\mathbf{m}_1, ..., \mathbf{m}_N\}$, where $\mathbf{m}_k \in \mathbb{R}^p$ is the vector of survey variables for the kth population unit, $k = 1, \ldots, N$. In the fixed-population, design-based approach to sampling, the values $\{\mathbf{m}_1, ..., \mathbf{m}_N\}$ are viewed as a collection of fixed, unknown constants. In order to estimate a population parameter $\theta = \theta(\mathbf{m})$, the values of the variables of interest are observed in a sample s of units selected from the population according to a probability distribution, $P(\cdot)$, that characterizes a sampling design, D. All expectations in this paper are taken with respect to $P(\cdot)$, unless we indicate the contrary. Let $\pi_k = P(u_k \in s)$ and $\pi_{kj} = P(u_k, u_j \in s)$ denote the first and second order inclusion probabilities, respectively. Let $I_k, k = 1, \ldots, N$, be the random variables defined as $I_k(s) = 1$ if $u_k \in s$ and $I_k(s) = 0$ otherwise, and let $\Delta_{kj} = \mathrm{Cov}(I_k, I_j)$, $k, j = 1, \ldots, N$. Along this paper we will assume that $\pi_k > 0, k = 1, \ldots, N$. Let $\widehat{T}_{HT}$ denote the Horvitz-Thompson (HT) estimator of the population total $T(\mathbf{m}) = \sum_{k=1}^{N} \mathbf{m}_k$,

$$\widehat{T}_{HT} = \sum_{k=1}^{N} \frac{\mathbf{m}_k}{\pi_k} I_k = \sum_{u_k \in s} \frac{\mathbf{m}_k}{\pi_k}.$$

By simplicity of notation, we denote $\sum_{k=1}^{N}$ by $\sum_U$ and $\sum_{u_k \in s}$ by $\sum_s$.

2 Regression in complex surveys

In model-based regression, the inference is based on a *model* that describes the relationship between the explanatory variables $X_1, ..., X_q$, and the response variable, Y. In particular, in the linear model-based regression, it is assumed that $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{Y}$ is an $n \times 1$ vector of observed values, $\mathbf{X}$ is an $n \times q$ matrix of known values for the n cases, β is a $q \times 1$ vector of unknown parameters and ε is an $n \times 1$ vector of i.i.d. random errors having mean zero and variance σ^2. In this framework, the ordinary least squares estimator of the parameter vector β is $\widehat{\beta} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$, the fitted values are defined by $\widehat{y}_i = \mathbf{x}_i^t\widehat{\beta}$, with $\mathbf{x}_i^t$ the ith row of $\mathbf{X}$ and $\mathrm{Var}_M(\widehat{\beta}) = \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$, where $\mathrm{Var}_M(\cdot)$ denotes the variance with respect to the assumed hypothesis in model-based regression.

Next, we consider the regression problem in complex survey. In this context, we will denote by $\mathbf{z}$ to the vector of explanatory variables. We assume that each population unit u_k, $k = 1, 2, \ldots, N$, has associated $q+1$ unknown characteristics, $\mathbf{m}_k^t = (y_k, \mathbf{z}_k^t)$, where $\mathbf{z}_k = \begin{pmatrix} z_{k1} & z_{k2} & \ldots & z_{kq} \end{pmatrix}^t$. The main objective of regression in complex surveys is to estimate the regression vector, $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \ldots & \beta_q \end{pmatrix}^t$, which is obtained through fitting the hyperplane $y = b_1 z_1 + \ldots + b_q z_q$ to the N population points by means of ordinary least squares, that is,

$$\beta = \arg\min_{\mathbf{b}\in\mathbb{R}^q} \sum_U (y_k - \mathbf{z}_k^t\mathbf{b})^2.$$

Note that we have denoted by β the regression parameter in both contexts, although its meaning is different in each case.

Let $\mathbb{E}_k = y_k - \mathbf{z}_k^t\beta$ denote the kth population residual. These residuals satisfy that $\sum_U \mathbf{z}_k\mathbb{E}_k = \mathbf{0}$, or equivalently, $\mathbf{T}\beta = \mathbf{t}$, where $\mathbf{T} = \sum_U \mathbf{z}_k\mathbf{z}_k^t$ and $\mathbf{t} = \sum_U \mathbf{z}_k y_k$. We assume that $\mathbf{T}$ is nonsingular and therefore $\beta = \mathbf{T}^{-1}\mathbf{t}$. Since both $\mathbf{T}$ and $\mathbf{t}$ are population totals, we can estimate them by their HT estimators,

$$\widehat{\mathbf{T}} = \sum_s \frac{1}{\pi_k}\mathbf{z}_k\mathbf{z}_k^t, \quad \widehat{\mathbf{t}} = \sum_s \frac{1}{\pi_k}\mathbf{z}_k y_k.$$

So, if $\widehat{\mathbf{T}}$ is nonsingular, we can consider the following estimator of β,

$$\widehat{\beta}_\pi = \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{t}}.$$

An alternative way to obtain $\widehat{\beta}_\pi$ is as follows: since $\beta = \arg\min_{\mathbf{b}} \tau(\mathbf{b})$, where $\tau(\mathbf{b}) = \sum_U (y_k - \mathbf{z}_k^t \mathbf{b})^2$ a population total, we can estimate β by means of $\widehat{\beta}_\pi^* = \arg\min_{\mathbf{b}} \widehat{\tau}(\mathbf{b})$, where $\widehat{\tau}(\mathbf{b}) = \sum_s (y_k - \mathbf{z}_k^t \mathbf{b})^2 \frac{1}{\pi_k}$ is the HT estimator of $\tau(\mathbf{b})$. Obviously, $\widehat{\beta}_\pi^* = \widehat{\beta}_\pi$ and therefore,

$$\sum_s \mathbf{z}_k (y_k - \mathbf{z}_k^t \widehat{\beta}_\pi) \frac{1}{\pi_k} = \sum_s \mathbf{z}_k e_k \frac{1}{\pi_k} = \mathbf{0},$$

where $e_k = y_k - \mathbf{z}_k^t \widehat{\beta}_\pi$ denotes the kth sample residual.

We remark that the properties of $\widehat{\beta}_\pi$ are determined by the probability distribution $P(\cdot)$, according to which the sample is selected, and those of $\widehat{\beta}$ are determined by the assumed hypothesis on the model. For example, $\widehat{\beta}$ is an unbiased estimator, but in general, $\widehat{\beta}_\pi$ is not an unbiased estimator, and their variances have very different expressions.

The influence study in a statistical analysis can include various aspects. We will center in the influence study on $\widehat{\beta}_\pi$.

3 Leverage and residuals

In model-based regression, the ith sample unit is said to be a high leverage point if the coefficient of y_i in the expression of $\widehat{y}_i = \mathbf{x}_i^t \widehat{\beta}$, given by $h_i = \mathbf{x}_i^t (\sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^t)^{-1} \mathbf{x}_i$, is large. In other words, the value of y_i dominates $\widehat{y}_i$. Geometrically this means that $\mathbf{x}_i$ is far from the rest of the $\mathbf{x}_j$ in the sample, i.e., $\mathbf{x}_i$ is an outlier in the space of the explanatory variables.

Analogously, in design-based regression one can look for those sample units having a large effect in its fitted value, $\widehat{y}_i = \mathbf{z}_i^t \widehat{\beta}_\pi$. If we denote by $\mathbf{Y}_s = (\ldots y_k \ldots)_{k \in s}^t$, we have that $\widehat{\mathbf{Y}}_s = \widetilde{\mathbf{H}}_s \mathbf{Y}_s$, where $\widetilde{\mathbf{H}}_s = (\widetilde{h}_{ij})_{i,j \in s}$, with $\widetilde{h}_{ij} = \mathbf{z}_i^t \left(\sum_s \mathbf{z}_k \mathbf{z}_k^t / \pi_k \right)^{-1} \mathbf{z}_j / \pi_j$. $\widetilde{\mathbf{H}}_s$ is an idempotent $n \times n$-matrix with rank q, but it is not a symmetric matrix, in general. This makes it hard to interpret those points with $\widetilde{h}_{ii}$ large. To facilitate its interpretation, remember that $\widehat{\beta}_\pi = \arg\min_{\mathbf{b}} \sum_s (y_k - \mathbf{z}_k^t \mathbf{b})^2 \frac{1}{\pi_k} = \arg\min_{\mathbf{b}} \sum_s (v_k - \mathbf{w}_k^t \mathbf{b})^2$, with $v_k = y_k / \pi_k^{1/2}$ and $\mathbf{w}_k = \mathbf{z}_k / \pi_k^{1/2}, \forall k \in s$. That is, the problem of least squares estimation of β in design-based regression is identical to the problem of least squares estimation of β in model-based regression, replacing y_i by v_i and $\mathbf{z}_i$ by $\mathbf{w}_i$. Hence, since $\widehat{\mathbf{V}}_s = \mathbf{P}_s \mathbf{V}_s$, where $\mathbf{P}_s = (p_{ij})$ with $p_{ij} = \mathbf{w}_i \left(\sum_s \mathbf{w}_k \mathbf{w}_k^t \right)^{-1} \mathbf{w}_j$, a unit in the sample, say u_i, is a high leverage point if p_{ii} is large. Note that $p_{ii} = \widetilde{h}_{ii}, \forall i \in s$. The advantage of considering $\mathbf{P}_s$ instead of $\widetilde{\mathbf{H}}_s$ is that $\mathbf{P}_s$ is a symmetric idempotent $n \times n$ matrix for any sampling design. This makes the geometric interpretation easy: a large $\widetilde{h}_{ii}$ means that $\mathbf{w}_i$ is far from the rest of the $\mathbf{w}_j$ in the sample, which implies that $\mathbf{z}_i$ is far from the rest of the $\mathbf{z}_j$ in the sample or that π_i is quite less than the rest of the π_j or both.

Hence, in design-based regression, the leverage points do not only depend on the relative position of the explanatory variables, as it occurs in model-based regression, but also on their first order inclusion probabilities. Trivially, if the sampling design is such that $\pi_i = c > 0$, the model-based leverage points and the design-based leverage points coincide.

In the model-based context different kind of residuals are employed to detect the presence of outliers with respect to regression. Basically, they are of two types: the ordinary residuals, which are the difference between y_i and $\widehat{y}_i$ and the standardized residuals, in which each ordinary residual is divided by a quantity that is proportional to its standard deviation. A sample point, say the ith sample point, with a large residual means that the behaviour of y_i in relation to $\mathbf{x}_i$ is quite different from the rest of the sample points.

Analogously, in design-based regression, one can examine the residuals: either the ordinary residuals, $e_i = y_i - \widehat{y}_i = y_i - \mathbf{z}_i^t \widehat{\beta}_\pi$, or some standardized residuals as

$$r_i = \frac{e_i}{\widehat{\mathrm{Var}}(e_i)^{1/2}}.$$

Since $\mathrm{Var}(e_i) = \mathbf{z}_i^t \mathrm{Var}(\widehat{\beta}_\pi)\mathbf{z}_i$, we will consider the following estimator of $\mathrm{Var}(e_i)$, $\widehat{\mathrm{Var}}(e_i) = \mathbf{z}_i^t \widehat{\mathrm{Var}}(\widehat{\beta}_\pi)\mathbf{z}_i$, and so we need an estimator of $\mathrm{Var}(\widehat{\beta}_\pi)$. From the first order Taylor's approximation for $\widehat{\beta}_\pi$ around β,

$$\widehat{\beta}_\pi \approx \widehat{\beta}_\pi^0 = \beta + \mathbf{T}^{-1}(\widehat{\mathbf{t}} - \widehat{\mathbf{T}}\beta), \tag{3.1}$$

the following approximation to the variance-covariance matrix of $\widehat{\beta}_\pi$ is obtained

$$\mathrm{Var}(\widehat{\beta}_\pi) \approx \mathrm{Var}(\widehat{\beta}_\pi^0) = \mathbf{T}^{-1}\mathbf{L}\mathbf{T}^{-1},$$

with

$$\mathbf{L} = \sum_{k=1}^{N}\sum_{j=1}^{N} \frac{\mathbb{E}_k \mathbb{E}_j}{\pi_k \pi_j}\, \Delta_{kj}\ \mathbf{z}_k \mathbf{z}_j^t.$$

Assuming $\pi_{kj} > 0, k, j = 1, \ldots, N$, we have the following estimator of $\mathrm{Var}(\widehat{\beta}_\pi)$, (e.g. Särndal et al., 1992, p. 194),

$$\widehat{\mathrm{Var}}(\widehat{\beta}_\pi) = \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{L}}\widehat{\mathbf{T}}^{-1},$$

where

$$\widehat{\mathbf{L}} = \sum_{u_k \in s} \sum_{u_j \in s} \frac{e_k e_j}{\pi_k \pi_j} \frac{\Delta_{kj}}{\pi_{kj}} \mathbf{z}_k \mathbf{z}_j^t.$$

Note that the expression of $\mathrm{Var}(\widehat{\beta}_\pi)$ is not as simple as $\mathrm{Var}_M(\widehat{\beta})$.

4 Influence diagnostics from conditional bias

4.1 Conditional bias in survey sampling: Estimation

Given a random sample $Y_1, ..., Y_n$ from a distribution function F, a statistic $R = R(Y_1, ..., Y_n)$ and a sample realization $y_1, ..., y_n$, Muñoz Pichardo et al. (1995) define the conditional bias of R given the ith observation as

$$E_F(R|Y_i = y_i) - E_F(R),$$

where E_F denotes the expectation with respect to the distribution F. By using this concept, Muñoz Pichardo et al. (1995) and Muñoz Pichardo et al. (2000) have obtained in a unified way a large number of influence measures which were previously proposed by several authors, each using a different argument.

In survey sampling, under design-based inference, in order to study the effect that the presence of the element u_i in the sample s has on the estimator $\widehat{\theta} = \widehat{\theta}(s)$, Moreno-Rebollo et al. (1999) define the conditional bias of $\widehat{\theta}$ due to the presence of u_i $(0 < \pi_i < 1)$ in the sample as

$$S(I_i = 1; \widehat{\theta}) = E(\widehat{\theta} \mid I_i = 1) - E(\widehat{\theta}).$$

$S(I_i = 1; \widehat{\theta})$ assesses the variation in the expected value of $\widehat{\theta}$ under a perturbation of the sampling design. The perturbation consists of restricting the sampling design to the samples containing u_i.

In general, $S(I_i = 1; \widehat{\theta})$ is a population parameter. Therefore, to obtain influence measures from it we must estimate it. As Moreno-Rebollo et al.

(1999) argue, the estimation should be carried out through the conditional sampling design, given the presence of u_i in the sample, $D_{|i}$, characterized by the probability function $P_{|i}(\cdot)$, defined by $P_{|i}(s) = \frac{P(s)}{\pi_i}$ if $u_i \in s$, $P_{|i}(s) = 0$, otherwise. The first order inclusion probabilities for $D_{|i}$ are given by $\pi_{k|i} = \pi_{ik}/\pi_i$.

In particular, the conditional bias of $\widehat{T}_{HT}$ is (Moreno-Rebollo et al., 1999)

$$S(I_i = 1; \widehat{T}_{HT}) = \sum_U \frac{\Delta_{ik}}{\pi_i \; \pi_k} \, \mathbf{m}_k. \tag{4.1}$$

Since $S(I_i = 1; \widehat{T}_{HT})$ is linear in the observed variables $\mathbf{m}_k$, it can be estimated by its HT estimator in $D_{|i}$,

$$\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT}) = \sum_s \frac{\Delta_{ik}}{\pi_{ik} \; \pi_k} \, \mathbf{m}_k.$$

Note that $\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT})$, like $S(I_i = 1; \widehat{T}_{HT})$, depends on the sampling design through the first and second order inclusion probabilities.

For any sampling design of fixed size, n, and for any estimator $\widehat{\theta}^{(n)}$, where the superscript (n) means that the estimator is designed for a sample of size n, Moreno-Rebollo et al. (2002) have proposed the following estimator of $S(I_i = 1; \widehat{\theta}^{(n)})$

$$\widehat{S}(I_i = 1; \widehat{\theta}^{(n)}) = (1 - \pi_i) \left\{ \widehat{\theta}^{(n)}(s) - \frac{P_{(i)}(s - \{u_i\})}{P_{|i}(s)} \widehat{\theta}^{(n-1)}(s - \{u_i\}) \right\}, \tag{4.2}$$

where $P_{(i)}(s - \{u_i\})$ is the probability of selecting the sample $s - \{u_i\}$, of size $n - 1$, in the design of fixed size $n - 1$ on $U_{(i)} = U - \{u_i\}$. This estimator can be seen as a finite population version of the sample influence curve (Cook and Weisberg, 1982, Chapter 3). Its application requires, besides the first and second order inclusion probabilities, knowledge of the probability distribution that defines the sampling design.

4.2 Conditional bias of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\pi}}$

In order to interpret $S(I_i = 1; \widehat{\beta}_\pi)$, to compare it with its counterpart in model-based regression and to propose an estimator, we will approximate

$S(I_i = 1; \widehat{\beta}_\pi)$ by $S(I_i = 1; \widehat{\beta}^0_\pi)$, with $\widehat{\beta}^0_\pi$ given by (3.1). Taking into account that

$$S(I_i = 1; \widehat{\beta}^0_\pi) = \mathbf{T}^{-1}\left[S(I_i = 1; \widehat{\mathbf{t}}) - S(I_i = 1; \widehat{\mathbf{T}})\beta\right], \tag{4.3}$$

and since from (4.1)

$$S(I_i = 1; \widehat{\mathbf{t}}) = \sum_U \frac{\Delta_{ik}}{\pi_i\ \pi_k}\, \mathbf{z}_k y_k \quad \text{and} \quad S(I_i = 1; \widehat{\mathbf{T}}) = \sum_U \frac{\Delta_{ik}}{\pi_i\ \pi_k}\, \mathbf{z}_k \mathbf{z}_k^t, \tag{4.4}$$

it is obtained that

$$S(I_i = 1; \widehat{\beta}_\pi) \approx S(I_i = 1; \widehat{\beta}^0_\pi) = \mathbf{T}^{-1}\left(\sum_U \frac{\Delta_{ik}}{\pi_i\ \pi_k}\, \mathbb{E}_k\ \mathbf{z}_k\right).$$

As $\sum_U \mathbb{E}_k \mathbf{z}_k = \mathbf{0}$, we also have the following expression for $S(I_i = 1; \widehat{\beta}^0_\pi)$

$$S(I_i = 1; \widehat{\beta}^0_\pi) = \mathbf{T}^{-1}\left(\sum_U \frac{\pi_{k|i}}{\pi_k} \mathbb{E}_k\ \mathbf{z}_k\right) = \frac{1}{\pi_i}\mathbb{E}_i \mathbf{T}^{-1}\mathbf{z}_i + \sum_{U-\{u_i\}} \frac{\pi_{k|i}}{\pi_k}\mathbb{E}_k \mathbf{T}^{-1}\mathbf{z}_k. \tag{4.5}$$

It is interesting to compare $S(I_i = 1; \widehat{\beta}^0_\pi)$ with its analogous in model-based regression, whose expression is (see Muñoz Pichardo et al., 1995)

$$\varepsilon_i \widetilde{\mathbf{T}}^{-1}\mathbf{x}_i, \tag{4.6}$$

where $\widetilde{\mathbf{T}} = \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^t$ and $\varepsilon_i = Y_i - \mathbf{x}_i^t \beta$. Comparing (4.5) and (4.6) we observe the following facts: (4.6) depends only on ε_i, whereas (4.5) is a linear combination of $\mathbb{E}_k \mathbf{T}^{-1}\mathbf{z}_k$, $k = 1, 2, ..., N$. The coefficients in this linear combination are determined by the first and second order inclusion probabilities associated with the sampling design. The term associated with u_i in the right-hand side of (4.5) is the analogue of (4.6) divided by π_i and therefore, this term increases as π_i decreases. The second term in the right-hand side of (4.5) can be interpreted as a consequence of the violation of the independence hypothesis in complex survey, where the presence of an observation in the sample may affect the inclusion probabilities of the remaining population units. On the coefficients of $\mathbb{E}_k \mathbf{T}^{-1}\mathbf{z}_k$ in this second term, $\frac{\pi_{k|i}}{\pi_k}$, we observe that:

- $\frac{\pi_{k|i}}{\pi_k}$ is null iff the sampling design is such that the elements u_k and u_i cannot appear simultaneously in the sample, that is, iff $\pi_{ik} = 0$.

- $\frac{\pi_{k|i}}{\pi_k}$ is > 1, $= 1$ or < 1 iff Δ_{ik} is > 0, $= 0$ or < 0, respectively, or equivalently, iff the first order inclusion probability of the element u_k is greater, equal or smaller, in the conditional design, given u_i, than that in the original one. In many sampling designs, $\Delta_{ik} < 0$ and in these designs $\mathbb{E}_i\mathbf{T}^{-1}\mathbf{z}_i$ dominates $S(I_i = 1; \widehat{\beta}^0_\pi)$ since it has the highest weight, because its coefficient is greater than 1 while the coefficient of $\mathbb{E}_k\mathbf{T}^{-1}\mathbf{z}_k$, $\forall k \neq i$, is less than 1.

As we have seen before, $S(I_i = 1; \widehat{\beta}^0_\pi)$ is a linear combination of $\mathbb{E}_k\mathbf{T}^{-1}\mathbf{z}_k$, $k = 1, 2, ..., N$, but there is some cases where it only depends on $\mathbb{E}_i\mathbf{T}^{-1}\mathbf{z}_i$. We say that a sampling design is independent if $\pi_{ik} = \pi_i\pi_k$, $\forall k \neq i$. Examples of independent designs are the Poisson and the Bernoulli designs. If a design is independent then

$$S(I_i = 1; \widehat{\beta}^0_\pi) = \frac{1 - \pi_i}{\pi_i}\mathbb{E}_i\,\mathbf{T}^{-1}\mathbf{z}_i.$$

Simple Random Sampling $(SRS(N, n))$ is not an independent design, but since it satisfies that $\frac{\Delta_{ik}}{\pi_k}$ is constant $\forall k \neq i$, we have that

$$S(I_i = 1; \widehat{\beta}^0_\pi) = \frac{N}{N-1}\frac{1-f}{f}\mathbb{E}_i\mathbf{T}^{-1}\mathbf{z}_i,$$

with $f = n/N$ the sampling fraction, that only depends on $\mathbb{E}_i\,\mathbf{T}^{-1}\mathbf{z}_i$.

4.3 Estimating the conditional bias

As we have previously indicated, the conditional bias is a population parameter and therefore, to obtain influence diagnostics from it we need an estimator. In this Section we will consider two estimators of $S(I_i = 1; \widehat{\beta}_\pi)$.

From the approximation (4.3) for $S(I_i = 1, \widehat{\beta}_\pi)$ and taking into account that the estimation should be carried out on the conditional sampling design, $D_{|i}$, we first consider the following estimator

$$\widetilde{S}(I_i = 1; \widehat{\beta}_\pi) = \widehat{\mathbf{T}}^{-1}_{|i}\left[\widehat{S}_{HT}(I_i = 1; \widehat{\mathbf{t}}) - \widehat{S}_{HT}(I_i = 1; \widehat{\mathbf{T}})\widehat{\beta}_{\pi|i}\right],$$

where $\widehat{\beta}_{\pi|i}$ is the least squares estimator of β in the conditional sampling design, that is, $\widehat{\beta}_{\pi|i} = \widehat{\mathbf{T}}^{-1}_{|i}\widehat{\mathbf{t}}_{|i}$. From (4.4) the following alternative expression is obtained

$$\widetilde{S}(I_i = 1; \widehat{\beta}_\pi) = \widehat{\mathbf{T}}_{|i}^{-1} \left(\sum_s \frac{\Delta_{ik}}{\pi_{ik}\ \pi_k}\, e_{k|i}\ \mathbf{z}_k \right),$$

where $e_{k|i} = y_k - \mathbf{z}_k^t \widehat{\beta}_{\pi|i} = y_k - \widehat{y}_{k|i}$. Taking into account that $\sum_s e_{k|i}\mathbf{z}_k \frac{1}{\pi_{k|i}} = \mathbf{0}$, we have that

$$\widetilde{S}(I_i = 1; \widehat{\beta}_\pi) = \widehat{\mathbf{T}}_{|i}^{-1} \left(\sum_s \frac{1}{\pi_k}\, e_{k|i}\ \mathbf{z}_k \right) \qquad (4.7)$$

$$= \widehat{\mathbf{T}}_{|i}^{-1}\widehat{\mathbf{T}} \left(\widehat{\beta}_\pi - \widehat{\beta}_{\pi|i} \right). \qquad (4.8)$$

Looking at (4.7) we conclude that $\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)$ is a linear combination of terms of the form $e_{k|i}\widehat{\mathbf{T}}_{|i}^{-1}\mathbf{z}_k$ where the coefficients are $\frac{1}{\pi_k}$. From (4.8) we observe that $\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)$ depends on the difference $\widehat{\beta}_\pi - \widehat{\beta}_{\pi|i}$, that is, of the difference between the estimators of β in the original design, D, and in the conditional design, given u_i, $D_{|i}$. This difference is analogous to that appearing in the case-deletion diagnostics, but while in the case-deletion diagnostics the original estimator is compared with the one obtained by omitting an observation, here we compare the original estimator with that obtained by imposing that $u_i \in s$.

Second, when the sampling design is of fixed size, n, we can consider the estimator proposed by Moreno-Rebollo et al. (2002) in (4.2), which in this case equals

$$\widehat{S}(I_i = 1; \widehat{\beta}_\pi) = (1 - \pi_i) \left\{ \widehat{\beta}_\pi(s) - \frac{P_{(i)}(s - \{u_i\})}{P_{|i}(s)} \widehat{\beta}_\pi^{(n-1)}(s - \{u_i\}) \right\},$$

with

$$\widehat{\beta}_\pi^{(n-1)}(s - \{u_i\}) = \left(\sum_{s-\{u_i\}} \frac{1}{\pi_k^{(n-1)}} \mathbf{z}_k \mathbf{z}_k^t \right)^{-1} \left(\sum_{s-\{u_i\}} \frac{1}{\pi_k^{(n-1)}} \mathbf{z}_k y_k \right),$$

where $\pi_k^{(n-1)}$ represents the first order inclusion probability in the design of fixed size $n - 1$ on U. Unlike $\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)$, $\widehat{S}(I_i = 1; \widehat{\beta}_\pi)$ can be viewed as a case-deletion diagnostic in survey sampling.

To conclude this subsection we give the expressions of the above proposed estimators in the $SRS(N,n)$:

$$\widetilde{S}(I_i=1;\widehat{\beta}_\pi)=\frac{N-n}{f(N-1)}e_{i|i}\widehat{\mathbf{T}}_{|i}^{-1}\mathbf{z}_i, \tag{4.9}$$

$$\widehat{S}(I_i=1;\widehat{\beta}_\pi)=(1-f)\left[\widehat{\beta}_\pi(s)-\widehat{\beta}_\pi^{(n-1)}(s-\{u_i\})\right]. \tag{4.10}$$

An alternative expression of (4.10) is given by

$$\widehat{S}(I_i=1;\widehat{\beta}_\pi)=\frac{1-f}{f}\frac{1}{1-\widetilde{h}_{ii}}e_i\widehat{\mathbf{T}}^{-1}\mathbf{z}_i.$$

Except for the finite population correction factor, $f/(1-f)$, $\widehat{S}(I_i=1;\widehat{\beta}_\pi)$ coincides with the estimator of the conditional bias of $\widehat{\beta}$, given $Y_i=y_i$, proposed by Muñoz Pichardo et al. (1995) in model-based regression, whose expression is

$$\frac{1}{1-h_{ii}}e_i^M\widetilde{\mathbf{T}}^{-1}\mathbf{x}_i=\widehat{\beta}-\widehat{\beta}_{(i)},$$

where $e_i^M=y_i-\mathbf{x}_i^t\widehat{\beta}$ and $\widehat{\beta}_{(i)}$ is the least squares estimate of β when the ith case is omitted from the study.

4.4 Influence diagnostics from the estimation of the conditional bias

Since $S(I_i=1;\widehat{\beta}_\pi)$, and its estimators, are q vectors, a way to obtain influence diagnostics is by normalizing the considered estimators, so that the observations can be ordered in a meaningful way. One of the methods to carry out this normalization, which is commonly used in model-based regression, is through a seminorm as (e.g. Chatterjee and Hadi, 1986)

$$\|\mathbf{x}\|^2_{\mathbf{M},c}=\frac{\mathbf{x}^t\mathbf{M}\mathbf{x}}{c}$$

for appropriate choices of $\mathbf{M}$ and c, where $\mathbf{M}$ is a positive semidefinite $q\times q-$matrix and $c\in\mathbb{R}$ is a positive constant. A large value of $\|\widetilde{S}(I_i=1;\widehat{\beta}_\pi)\|_{\mathbf{M},c}$ indicates that the ith observation has a high influence on $\widehat{\beta}_\pi$ relative to $\mathbf{M}$ and c. A similar interpretation applies to $\|\widehat{S}(I_i=1;\widehat{\beta}_\pi)\|_{\mathbf{M},c}$.

Several influence diagnostics have been proposed in model-based regression normalizing $\widehat{\beta} - \widehat{\beta}_{(i)}$, where $\widehat{\beta}_{(i)}$ is the least square estimator of β obtained by omitting the ith observation,

$$\widehat{\beta}_{(i)} = \arg\min_{\mathbf{b}} \sum_{k \neq i} (y_k - \mathbf{x}_k^t \mathbf{b})^2,$$

considering $\mathbf{M} = \mathbf{X}^t\mathbf{X} = \widetilde{\mathbf{T}}$, or the analogue of this matrix in the reduced sample, and c an estimator of σ^2 from the whole sample or under the omission. A justification to this choice of $\mathbf{M}$ and c is found in the fact that $\mathrm{Var}_M(\widehat{\beta}) = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1} = \sigma^2\widetilde{\mathbf{T}}^{-1}$.

By analogy, we can define influence diagnostics in design-based regression following this procedure by considering for example $\mathbf{M} = \{\widehat{\mathrm{Var}}(\widehat{\beta}_\pi)\}^{-1}$ $= \widehat{\mathbf{T}}\widehat{\mathbf{L}}^{-1}\widehat{\mathbf{T}}$ and $c = 1$ or $\mathbf{M} = \widehat{\mathbf{T}}$ and $c = c_1$ or $c = c_2$, where

$$c_1 = \frac{1}{n-q}\sum_s \frac{e_k^2}{\pi_k} \quad \text{and} \quad c_2 = \frac{1}{n-q}\sum_s \frac{e_{k|i}^2}{\pi_{k|i}}.$$

The reason for these choices of the constant c is that, as said before, in model-based regression a common choice of c is an estimator of σ^2. Although the variance σ^2 does not have an analogous in design-based regression, we can construct analogues of some of its estimators as follows. An unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-q} SC_\varepsilon,$$

where

$$SC_\varepsilon = \sum_{k=1}^{n} (y_k - \mathbf{x}_k^t \widehat{\beta})^2 = \min_{\mathbf{b}} \sum_{k=1}^{n} (y_k - \mathbf{x}_k^t \mathbf{b})^2.$$

Taking into account that $\widehat{\beta}_\pi = \arg\min_{\mathbf{b}} \sum_s (y_k - \mathbf{z}_k^t \mathbf{b})^2 \frac{1}{\pi_k}$, the analogue of SC_ε in design-based regression can be taken SC_e, where

$$SC_e = \min_{\mathbf{b}} \sum_s (y_k - \mathbf{z}_k^t \mathbf{b})^2 \frac{1}{\pi_k} = \sum_s \frac{e_k^2}{\pi_k},$$

which justifies the choice of $c = c_1$. Another usual choice of c in model-based regression is

$$\widehat{\sigma}_{(i)}^2 = \frac{1}{n-q} SC_{\varepsilon(i)},$$

where

$$SC_{\varepsilon(i)} = \sum_{k=1}^{n} (y_k - \mathbf{x}_k^t \widehat{\beta}_{(i)})^2.$$

If instead of considering the omission we consider the conditional design $D_{|i}$ (recall the discussion on expression (4.8)) and reasoning as before we obtain c_2.

Other choices of $\mathbf{M}$ are possible. Table 1 displays several influence diagnostics obtained for different choices of $\mathbf{M}$ and $c = 1$.

Table 1: Influence diagnostics obtained for different choices of $\mathbf{M}$ *and* $c = 1$

$\mathbf{M}$	$\Vert\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\Vert^2_{\mathbf{M},1}$
$\mathbf{M}_1 = \widehat{\mathbf{T}}_{\vert i}\widehat{\mathbf{T}}^{-1}\widehat{\mathbf{T}}_{\vert i}$	$\sum_s \frac{\left(\widehat{y}_k - \widehat{y}_{k\vert i}\right)^2}{\pi_k} = \sum_s \frac{\left(e_k - e_{k\vert i}\right)^2}{\pi_k}$
$\mathbf{M}_2 = \widehat{\mathbf{T}}_{\vert i}$	$\sum_{\substack{u_k \in s \\ u_j \in s}} \frac{\left(\widehat{y}_k - \widehat{y}_{k\vert i}\right)\left(\widehat{y}_j - \widehat{y}_{j\vert i}\right)}{\pi_k \pi_j} \widetilde{h}_{kj\vert i}\pi_{j\vert i}$
$\mathbf{M}_3 = \widehat{\mathbf{T}}$	$\sum_{\substack{u_k \in s \\ u_j \in s}} \frac{\left(\widehat{y}_k - \widehat{y}_{k\vert i}\right)\left(\widehat{y}_j - \widehat{y}_{j\vert i}\right)}{\pi_k \pi_j} \mathbf{z}_k^t \widehat{\mathbf{T}}_{\vert i}^{-1} \widehat{\mathbf{T}} \widehat{\mathbf{T}}_{\vert i}^{-1} \mathbf{z}_j$
$\mathbf{M}_4 = \left\{\widehat{\mathrm{Var}}\left(\widehat{\beta}_\pi\right)\right\}^{-1}$	$\sum_{\substack{u_k \in s \\ u_j \in s}} \frac{\left(\widehat{y}_k - \widehat{y}_{k\vert i}\right)\left(\widehat{y}_j - \widehat{y}_{j\vert i}\right)}{\pi_k \pi_j} \mathbf{z}_k^t \widehat{\mathbf{T}}_{\vert i}^{-1} \widehat{\mathbf{T}} \widehat{\mathbf{L}}^{-1} \widehat{\mathbf{T}} \widehat{\mathbf{T}}_{\vert i}^{-1} \mathbf{z}_j$

For the choices of $\mathbf{M}$ in Table 1, all the obtained influence diagnostics depend on the differences between the fitted values in the original design and in the conditional design, $\widehat{y}_k - \widehat{y}_{k|i}$, or equivalently, on the residuals, $e_k - e_{k|i}$, on the elements of the matrix $\widetilde{\mathbf{H}}$ and on the first and second order inclusion probabilities. In general, the expressions of the resultant diagnostics are more complex than the ones obtained in model-based regression. This is a consequence of the relations of dependence among the observations in each context.

The diagnostic $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{\mathbf{M}_1, c_1}$ is formally the simplest one and it is interesting to compare it with Cook's distance, proposed by Cook (1977) as an influence diagnostic in model-based regression, given by

$$\frac{\left(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)}\right)^t \left(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)}\right)}{q\widehat{\sigma}^2}. \tag{4.11}$$

Cook's distance is proportional to the euclidean norm of $(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})$, while $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{\mathbf{M}_1, c_1}$ is proportional to the norm of $(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{|i})$ with respect to the diagonal matrix diag$\{\ldots \pi_k^{-1} \ldots\}$, that is, while the diagnostic (4.11) gives the same importance to all observations, in $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{\mathbf{M}_1, c_1}$ the weight of each observation is inversely proportional to its first order inclusion probability.

Analogously, influence measures can be defined from $\widehat{S}(I_i = 1; \widehat{\beta}_\pi)$.

4.5 SRS and independent designs

The generality of the expressions in Table 1 makes hard its interpretation. In this subsection we evaluate them in some particular designs: the $SRS(N, n)$ and independent designs. Each of them has a characteristic in common with the i.i.d. case: in a SRS the variables I_i are identically distributed and in the case of an independent design they are independent.

The results in Table 2 are obtained from the expressions of $\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)$ in a SRS and in the independent designs, given by (4.9) for a SRS and by

$$\widetilde{S}(I_i = 1; \widehat{\beta}_\pi) = \frac{1 - \pi_i}{\pi_i} e_{i|i}\, \widehat{\mathbf{T}}_{|i}^{-1} \mathbf{z}_i$$

for independent designs, and taking into account that in a SRS: $e_{i|i} = \frac{1}{1-g_{ii}} e_i$, where $g_{ii} = \frac{N-n}{N-1}\widetilde{h}_{ii}$, and $\widehat{\mathbf{T}}_{|i}^{-1} z_i = \frac{n-1}{f(N-1)} \frac{1}{1-g_{ii}} \widehat{\mathbf{T}}^{-1} z_i$, and that for independent designs: $e_{i|i} = \frac{1}{1-\tau_{ii}} e_i$, where $\tau_{ii} = (1 - \pi_i)\widetilde{h}_{ii}$ and $\widehat{\mathbf{T}}_{|i}^{-1} z_i = \frac{1}{1-\tau_{ii}} \widehat{\mathbf{T}}^{-1} z_i$.

In the case of a SRS the diagnostics associated with the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, except for a constant factor, can be expressed as a product of the form $d_j(g_{ii}) e_i^2$, $j = 1, 2, 3$, with $d_j(x) = \frac{x}{(1-x)^{j+1}}$. The functions $d_j(g_{ii})$, $j = 1, 2, 3$, are increasing in g_{ii} since $g_{ii} \in [0, 1]$ and they also

Table 2: Influence diagnostics in $SRS(N,n)$ and independent designs

	$\|\widetilde{S}(I_i=1;\widehat{\beta}_\pi)\|^2_{\mathbf{M},1}$	
	SRS	Independent
$\mathbf{M}_1$	$\frac{(N-n)}{(N-1)f}\frac{g_{ii}}{(1-g_{ii})^2}e_i^2$	$\frac{1-\pi_i}{\pi_i}\frac{\tau_{ii}}{(1-\tau_{ii})^2}e_i^2$
$\mathbf{M}_2$	$\frac{(N-n)(n-1)}{(N-1)^2f^2}\frac{g_{ii}}{(1-g_{ii})^3}e_i^2$	$\frac{1-\pi_i}{\pi_i}\frac{\tau_{ii}}{(1-\tau_{ii})^3}e_i^2$
$\mathbf{M}_3$	$\frac{(N-n)(n-1)^2}{(N-1)^3f^3}\frac{g_{ii}}{(1-g_{ii})^4}e_i^2$	$\frac{1-\pi_i}{\pi_i}\frac{\tau_{ii}}{(1-\tau_{ii})^4}e_i^2$
$\mathbf{M}_4$	$\frac{(N-n)^2}{(N-1)^2f^2}\frac{1}{(1-g_{ii})^2}e_i^2\,\mathbf{z}_i^t\widehat{\mathbf{L}}_{\|i}^{-1}\mathbf{z}_i$	$\frac{(1-\pi_i)^2}{\pi_i^2}\frac{1}{(1-\tau_{ii})^2}e_i^2\,\mathbf{z}_i^t\widehat{\mathbf{L}}_{\|i}^{-1}\mathbf{z}_i$

satisfy that $d_1 \leq d_2 \leq d_3$. Therefore, we can conclude that in these three cases the influence diagnostics are multiplicative functions of a measure of leverage and of the residual e_i, and they differ in the impact of the leverage on the diagnostic, since $d_1 \leq d_2 \leq d_3$.

In the case of an independent design the influence measures obtained from $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, can be expressed as $\frac{1-\pi_i}{\pi_i}d_j(\tau_{ii})e_i^2$. The same comment as before can be done on $d_j(\tau_{ii}), j=1,2,3$, since $\tau_{ii}\in[0,1]$. Therefore, in this case the influence diagnostics associated with $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$ are multiplicative functions of a measure of leverage, the residual e_i and of a decreasing function in π_i, $\frac{1-\pi_i}{\pi_i}$.

5 Examples

We present two examples to illustrate the proposed diagnostics. The first one is an artificial bivariate data set, where it is easy to identify influential cases, from the model-based point of view, in a scatter plot. This will allow us to observe the differences between model-based diagnostics and the design-based diagnostics proposed in this paper. The second example is a real data set.

5.1 Artificial data set

We have generated an artificial population of size $N = 250$. The vector of survey variables is $\mathbf{m}_k = (y_k, 1, z_k)^t$, $k = 1, ..., N$. To estimate β, the regression parameter of y on the explanatory variables, we have selected a sample of size $n = 50$ from the population according to an inclusion probability proportional to size design, constructed by Sampford's method, (Sampford, 1967). The sample data are given in Table 3.

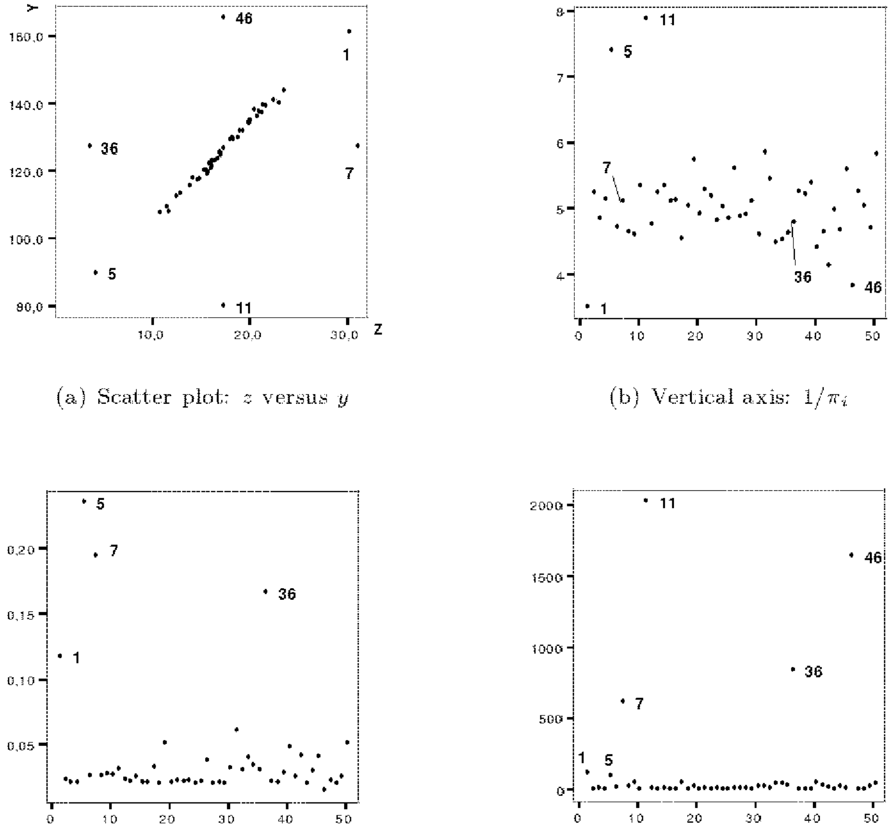
Table 3: Artificial data set: (y) response variable, (z) explanatory variable, (π_i^{-1}) inverse of the first order inclusion probability

Case	y	z	π_i^{-1}	Case	y	z	π_i^{-1}
1	161,01	30,01	3,50	26	113,06	12,58	5,61
2	120,07	15,02	5,25	27	128,96	17,82	4,88
3	129,94	17,95	4,85	28	129,58	17,92	4,90
4	122,68	15,89	5,14	29	123,7	16,45	5,09
5	89,72	3,89	7,39	30	137,35	21,05	4,58
6	134,11	19,71	4,70	31	107,51	10,59	5,85
7	127	30,95	5,11	32	115,73	13,62	5,44
8	134,86	19,85	4,64	33	141	22,22	4,47
9	137,99	20,23	4,58	34	139,38	21,44	4,52
10	117,35	14,35	5,35	35	137,45	20,72	4,61
11	79,92	17,06	7,87	36	127,05	3,25	4,78
12	131,67	19,07	4,75	37	120,91	15,7	5,25
13	119,66	15,46	5,24	38	121,4	15,9	5,20
14	117,62	14,6	5,35	39	117,82	13,88	5,39
15	123,17	16,19	5,10	40	143,95	23,16	4,40
16	123,24	15,85	5,13	41	135	19,75	4,63
17	139,54	21,18	4,53	42	140,01	22,79	4,13
18	124,64	16,67	5,03	43	126,48	17,08	4,98
19	109,42	11,3	5,74	44	135,96	20,54	4,66
20	129,09	18,03	4,91	45	112,31	12,26	5,59
21	119,14	15,34	5,28	46	165,41	17,06	3,82
22	122,25	15,56	5,17	47	120,02	15,21	5,25

(*Continued on next page*)

(Table 3. Continued from previous page)

Case	y	z	π_i^{-1}	Case	y	z	π_i^{-1}
23	131,53	18,73	4,81	48	124,88	16,77	5,03
24	125,51	16,64	5,01	49	134,27	19,62	4,69
25	129,71	18,48	4,85	50	107,78	11,44	5,82



(a) Scatter plot: z versus y

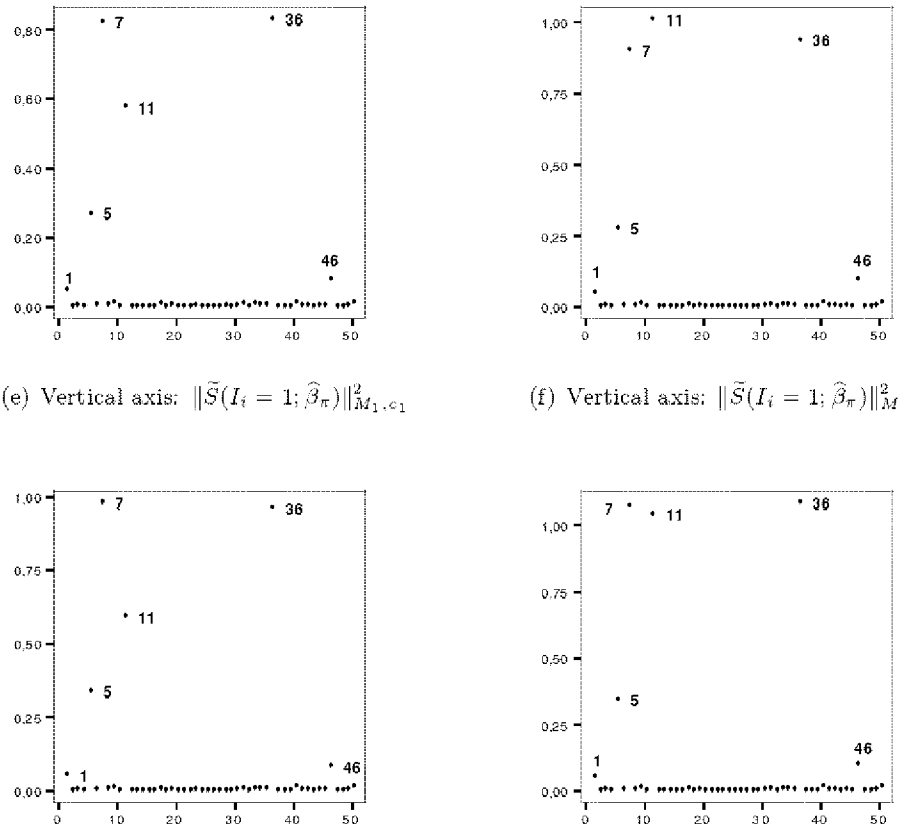
(b) Vertical axis: $1/\pi_i$

(c) Vertical axis: $\widetilde{h}_{ii}$

(d) Vertical axis: e_i^2

Figure 1: Artificial data set (in graphics 1(b)–1(d), horizontal axis: cases)

The auxiliary variable is approximately proportional to y, and so from Figure 1(a) we can get an idea of the first order inclusion probabilities of



(e) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_1, c_1}$

(f) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_1, c_2}$

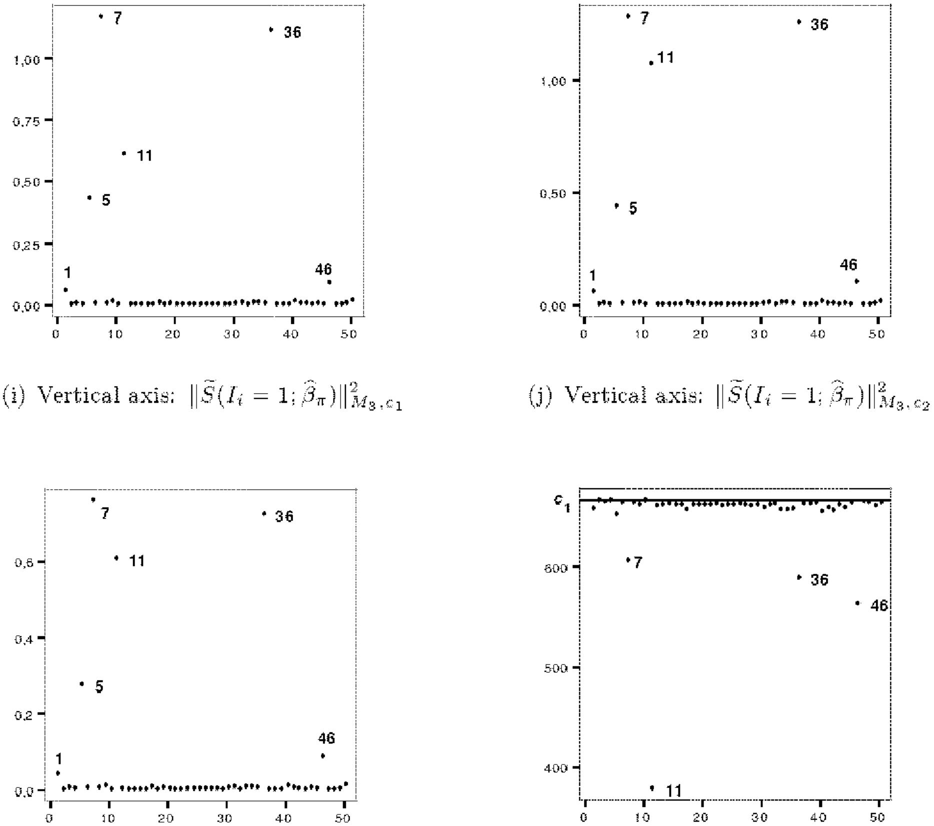
(g) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_2, c_1}$

(h) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_2, c_2}$

Figure 1 (continuation): Artificial data set (horizontal axis: cases)

the sample units. Nevertheless, the values of $1/\pi_k$, $k \in s$, are displayed in Figure 1(b).

Looking at Figure 1(a) we see there are six cases that, from the model-based point of view, should be considered as influential: cases 1 and 5 because they are leverage; cases 11 and 46 because they have a large residual; and cases 7 and 36 because they are leverage and they also have large residuals.



(i) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_3,c_1}$

(j) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_3,c_2}$

(k) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_4,1}$

(l) Vertical axis: c_2

Figure 1 (continuation): Artificial data set (horizontal axis: cases)

Next we examine the data from the design-based point of view. Figure 1(c) displays the values $\widetilde{h}_{ii}$. Looking at it we can clearly see the effect of the design: although in model-based context cases 1 and 5 both have almost the same h_{ii}, here case 5 is more leverage than case 1. This difference is due to the fact that π_1 is much bigger than π_5, as it can be seen from Figure 1(b). Cases 7 and 36 have similar $\widetilde{h}_{ii}$ because they have similar π_i. Figure 1(d) shows e_i^2, $i \in s$. We observe that the relative positions of e_i^2, $i \in s$, are quite similar to those that would be obtained in model-based regression.

Figures 1(e)-1(k) display the values of $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{\mathbf{M},c}$ for $(\mathbf{M}, c) = (\mathbf{M}_1, c_1)$, $(\mathbf{M}_1, c_2)$, $(\mathbf{M}_2, c_1)$, $(\mathbf{M}_2, c_2)$, $(\mathbf{M}_3, c_1)$, $(\mathbf{M}_3, c_2)$, $(\mathbf{M}_4, 1)$, respectively. All the diagnostics declare cases 7 and 36 as highly influential, 5 as a moderate influential case and cases 1 and 46 as low influential. Case 11 is classified as high influential by some measures and as moderate influential by others. Looking at Figures 1(e)-1(j) we observe that the relative position of case 11 depends on the constant being considered, whose values are shown in Figure 1(l).

5.2 Real data set

In this subsection we present an example with real data. The considered population consists of those towns of Andalucía (a region in the south of Spain) with a number of residents (R) in the interval $[200, 25000]$ and with electric energy consumption (EC) less than 150000 megawatts per hour. The population size is $N = 708$. The response variable (y) is the declared net rents in the $IRPF$ tax and the explanatory variables are EC and the number of enterprises (NE). All data are referred to 1999 and they have been obtained in June 2003 from `www.juntadeandalucia.es/institutodeestadistica`.

We have selected a sample according to a proportional probability aggregated size sampling (PPAS) (Hedayat and Sinha, 1991, Chapter 6), with sample size $n = 120$ and size variable R. The sample data are given in Table 4.

Table 4: Real data set: (IRPF) declared rent in the IRPF tax, (EC) electric energy consum, (NE) number of enterprises, (π_i^{-1}) inverse of the first order inclusion probability

Case	IRPF	NE	EC	π_i^{-1}	Case	IRPF	NE	EC	π_i^{-1}
1	46834,23	869	41,98	5,78	61	43914,55	1140	49,48	5,77
2	3465,56	86	2,47	5,93	62	543,71	15	0,36	5,94
3	23352,40	870	15,73	5,85	63	5982,06	106	5,62	5,92
4	401,15	7	0,21	5,94	64	19514,45	215	89,11	5,89
5	8040,56	217	5,74	5,91	65	7451,36	157	7,29	5,90
6	283,11	22	0,57	5,94	66	309,99	9	0,31	5,94
7	1324,23	49	1,29	5,93	67	3130,08	129	2,71	5,92
8	45624,34	1071	35,61	5,81	68	4356,83	167	2,53	5,92

(*Continued on next page*)

(*Table 4. Continued from previous page*)

Case	IRPF	NE	EC	π_i^{-1}	Case	IRPF	NE	EC	π_i^{-1}
9	740,48	27	1,66	5,94	69	19766,19	429	10,90	5,87
10	969,55	26	0,66	5,94	70	550,74	41	0,40	5,94
11	16195,34	532	18,27	5,88	71	1419,41	39	1,53	5,93
12	587,47	16	0,75	5,94	72	3100,63	73	1,79	5,93
13	13770,91	638	17,80	5,89	73	5124,90	105	4,86	5,91
14	538,35	13	0,30	5,94	74	4908,71	130	4,72	5,91
15	1403,62	63	0,93	5,93	75	7137,12	100	7,51	5,92
16	543,12	16	0,56	5,94	76	1400,73	51	1,28	5,93
17	4588,40	113	3,64	5,92	77	11886,29	232	6,35	5,89
18	6405,24	175	6,17	5,89	78	1326,76	23	3,45	5,93
19	2867,04	82	4,60	5,92	79	33510,94	730	37,78	5,85
20	4353,15	131	3,79	5,92	80	20378,80	411	17,23	5,87
21	13924,62	320	7,86	5,88	81	23885,65	570	90,29	5,86
22	7198,29	153	12,93	5,90	82	3989,03	79	3,14	5,92
23	6059,60	127	4,69	5,92	83	22439,52	452	12,94	5,87
24	6791,25	223	5,71	5,90	84	4806,47	154	2,99	5,92
25	10411,80	292	13,66	5,90	85	13972,06	285	9,78	5,88
26	1223,17	52	0,81	5,93	86	19243,33	436	14,57	5,87
27	6834,40	167	6,95	5,92	87	2026,98	43	1,21	5,93
28	26517,41	600	14,36	5,85	88	8489,40	210	25,77	5,91
29	1052,95	31	1,02	5,94	89	13366,76	244	8,74	5,89
30	12300,25	273	7,21	5,89	90	7056,72	327	5,33	5,90
31	12196,74	323	1,47	5,88	91	588,59	25	0,46	5,94
32	19514,96	463	19,68	5,87	92	2308,24	52	1,56	5,92
33	9299,04	207	10,03	5,91	93	4218,62	212	4,19	5,92
34	1321,09	36	1,26	5,93	94	3815,19	96	2,64	5,92
35	22569,06	594	15,87	5,85	95	1364,71	36	0,66	5,93
36	1194,76	18	0,85	5,93	96	306,89	18	0,34	5,94
37	12680,60	246	12,01	5,90	97	1592,17	62	1,94	5,93
38	2115,23	64	1,65	5,92	98	333,48	17	0,33	5,94
39	3846,65	68	3,32	5,93	99	1790,37	35	0,80	5,93
40	2436,69	46	1,97	5,93	100	1073,20	41	0,70	5,93
41	3698,84	47	2,38	5,93	101	17781,31	509	20,93	5,89
42	1327,09	39	1,46	5,94	102	4023,17	105	2,28	5,92
43	511,02	27	0,41	5,94	103	6669,54	201	11,41	5,91

(*Continued on next page*)

(Table 4. Continued from previous page)

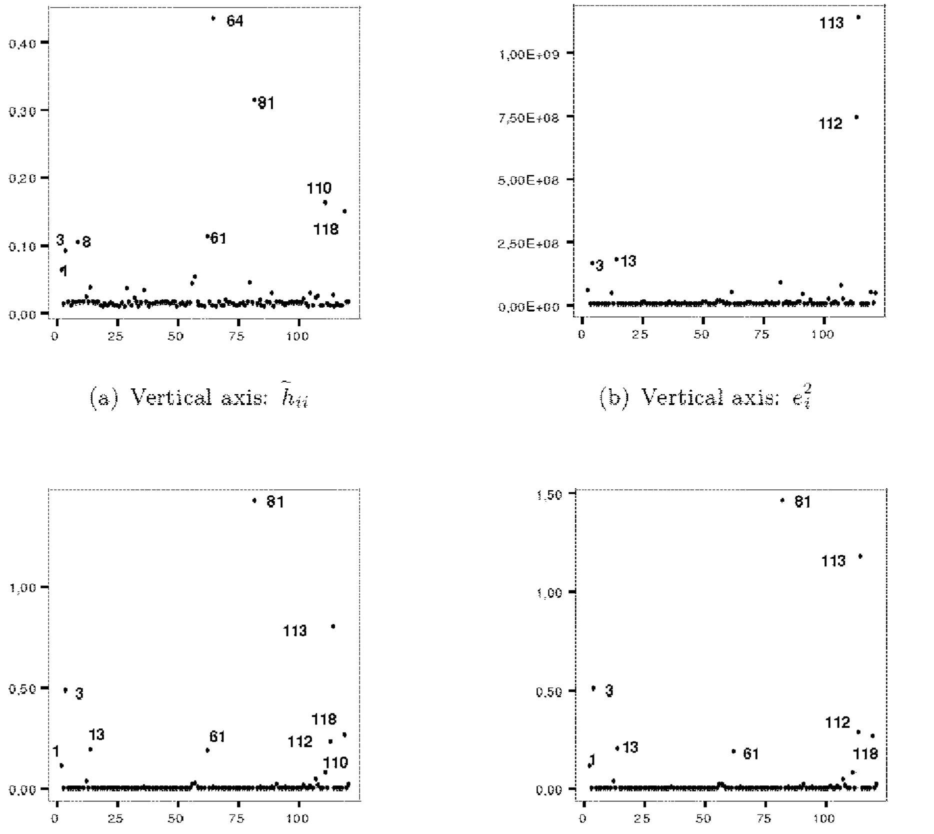
Case	IRPF	NE	EC	π_i^{-1}	Case	IRPF	NE	EC	π_i^{-1}
44	20034,16	463	18,98	5,84	104	3918,81	126	22,39	5,91
45	1167,20	31	2,94	5,93	105	9830,35	254	6,74	5,90
46	1719,16	73	1,48	5,93	106	29960,62	499	15,50	5,82
47	439,13	15	0,34	5,94	107	17764,23	526	15,35	5,86
48	1149,18	15	1,57	5,94	108	4683,51	125	3,28	5,92
49	13892,28	266	9,02	5,87	109	11784,68	281	9,11	5,89
50	1245,35	101	2,83	5,93	110	59513,90	1292	40,89	5,72
51	1326,48	55	1,13	5,93	111	3437,31	80	1,92	5,93
52	5071,14	133	3,00	5,92	112	37369,30	216	15,27	5,89
53	1030,14	52	0,98	5,93	113	57676,56	555	18,26	5,85
54	1288,60	37	0,99	5,93	114	10523,87	253	12,27	5,90
55	2880,97	94	26,13	5,93	115	4334,25	91	3,09	5,92
56	38683,80	763	41,87	5,83	116	4868,20	146	3,80	5,92
57	11546,40	347	6,69	5,89	117	7148,44	163	4,47	5,91
58	2985,25	100	3,82	5,92	118	56046,67	1016	74,24	5,75
59	3215,96	140	2,25	5,93	119	16947,18	427	21,94	5,86
60	7573,28	164	6,87	5,91	120	10837,65	410	13,74	5,88

Figures 2(a) and 2(b) display $\widetilde{h}_{ii}$ and e_i^2, $i \in s$, respectively. Looking at these figures we conclude that cases 64 and 81 are high leverage points and cases 1, 3, 8, 61, 110 and 118 have moderate leverage; cases 112 and 113 have large residuals and cases 3, 13, 61, 81 and 106 have moderate residuals.

Figures 2(c)–2(i) display the values of $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{\mathbf{M},c}$ for $(\mathbf{M}, c) = (\mathbf{M}_1, c_1)$, $(\mathbf{M}_1, c_2)$, $(\mathbf{M}_2, c_1)$, $(\mathbf{M}_2, c_2)$, $(\mathbf{M}_3, c_1)$, $(\mathbf{M}_3, c_2)$, $(\mathbf{M}_4, 1)$, respectively. All the considered diagnostics declare case 81 as high influential. Looking at Figures 2(c)–2(h) we see that when $c = c_2$ the influence of case 113 increases. This is due to the value of c_2 for this case, as can be seen from Figure 2(j). As in the artificial data set example, the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$ give similar results. The matrix $\mathbf{M}_4$ only changes the relative position of case 113.

For a PPAS sampling we have that

$$\frac{P_{(i)}(s-u_i)}{P_{|i}(s)} = \frac{t_s(R)-R_i}{(n-1)t_s(R)}\left\{(n-1)+\frac{(N-1)R_i}{T(R)-R_i}\right\}, \qquad (5.1)$$



(a) Vertical axis: $\widetilde{h}_{ii}$

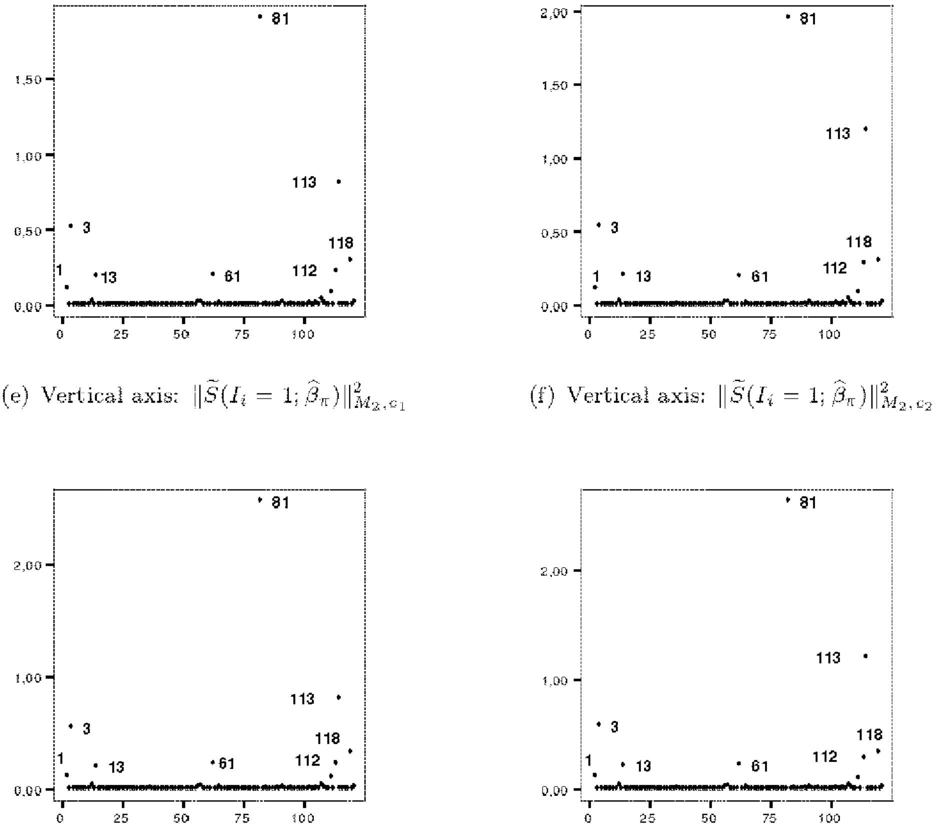
(b) Vertical axis: e_i^2

(c) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_1,c_1}$

(d) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_1,c_2}$

Figure 2: Real data set (in all graphics, horizontal axis: cases)

where $t_s(R)$ is the sample total and $T(R)$ is the population total. Since for this sampling (5.1) is known, in this example we can also compute $\widehat{S}(I_i = 1; \widehat{\beta}_\pi)$ and obtain influence measures from it. We have calculated $\|\widehat{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{\mathbf{M},c}$ for the values of $(\mathbf{M}, c)$ previously considered. Comparing $\|\widehat{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{\mathbf{M},c}$ with $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{\mathbf{M},c}$ we observe that the influence diagnostics obtained from $\widehat{S}(I_i = 1; \widehat{\beta}_\pi)$ increase the relative position of case 113, for all the considered choices of $(\mathbf{M}, c)$. To illustrate this, see Figures 2(c) and 2(k).



(e) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_2,c_1}$

(f) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_2,c_2}$

(g) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_3,c_1}$

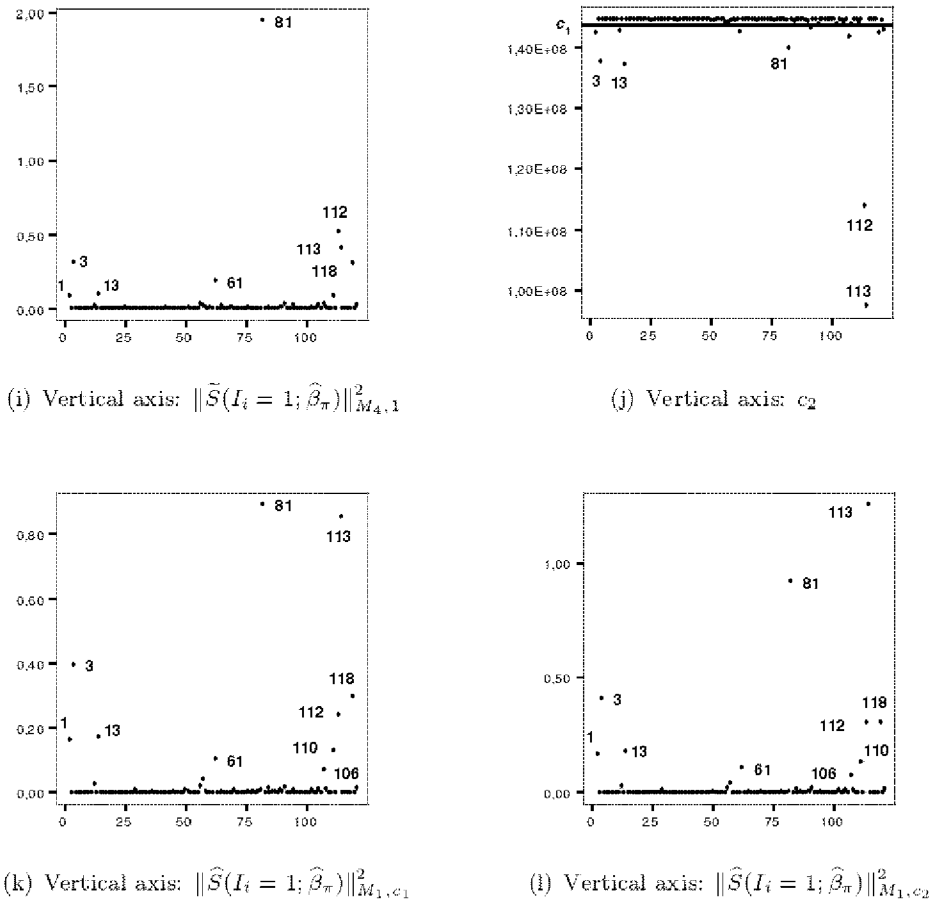
(h) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_3,c_2}$

Figure 2 (continuation): Real data set

6 Conclusions

In this paper we have defined some influence measures in design-based regression, area in which the influence analysis has been scarcely treated.

First, we introduce the analogues of leverage and residuals for design-based regression. In particular, we observe that in design-based regression, the leverage points do not only depend on the relative position of the ex-



(i) Vertical axis: $\|\widetilde{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_4,1}$

(j) Vertical axis: c_2

(k) Vertical axis: $\|\widehat{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_1,c_1}$

(l) Vertical axis: $\|\widehat{S}(I_i = 1; \widehat{\beta}_\pi)\|^2_{M_1,c_2}$

Figure 2 (continuation): Real data set

planatory variables, as it occurs in model-based regression, but also on their first order inclusion probabilities.

The diagnostics that we propose are built by normalizing the conditional bias (Moreno-Rebollo et al., 1999) of $\widehat{\beta}_\pi$, the usual estimator of the regression parameter in design-based regression. The resultant diagnostics are compared with their counterpart in model-based regression. From this comparison we observe that the proposed diagnostics depend on the residuals and leverage, like most model-based regression influence measures. However, in design-based regression, the diagnostics also depend on the first and second order probabilities associated with the sampling design. That is, the proposed diagnostics incorporate probability weighting, population structure or measures of size into the analysis. The results obtained show clearly what Smith (1987) asserts "conventional model-based influence diagnostics do not have immediate application to randomization inference for sample survey".

Acknowledgement

The authors thank the referee and the associate editor for their detailed comments and constructive suggestions.

References

BARRETT, B. E. and LING, R. F. (1992). General classes of influence measures for multivariate regression. *Journal of the American Statistical Association*, 87:184–191.

BELSLEY, D. A., KUH, E., and WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, New York.

CARONI, C. (1987). Residuals and influence in the multivariate linear model. *The Statistician*, 36:365–370.

CHATTERJEE, S. and HADI, A. S. (1986). Influential observations, high leverage points and outliers in linear regression. *Statistical Science*, 1(3):379–416.

CHATTERJEE, S. and HADI, A. S. (1988). *Sensitivity Analysis in Linear Regression*. John Wiley & Sons, New York.

COOK, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19:15–18.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 55(2):193–203.

Gwet, J. P. and Rivest, L. P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of American Statistical Association*, 87(420):79–87.

Hedayat, A. S. and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling*. John Wiley & Sons, New York.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21(1):79–87.

Moreno-Rebollo, J. L., Muñoz Reyes, A., Jiménez-Gamero, M. D., and Muñoz Pichardo, J. M. (2002). Influence diagnostics in survey sampling: Estimating the conditional bias. *Metrika*, 55:209–214.

Moreno-Rebollo, J. L., Muñoz Reyes, A., and Muñoz Pichardo, J. M. (1999). Influence diagnostics in survey sampling: Conditional bias. *Biometrika*, 84(4):923–928.

Muñoz Pichardo, J. M., Muñoz García, J., Fernández-Ponce, J. M., and Jiménez-Gamero, M. D. (2000). Influence analysis in multivariate linear general models. *Communications in Statistics. Theory and Methods*, 29:529–547.

Muñoz Pichardo, J. M., Muñoz García, J., Moreno-Rebollo, J. L., and Pino-Mejías, R. (1995). A new approach to influence analysis in linear models. *Sankhyā. Series A*, 57:393–409.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54:499–513.

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

Smith, T. M. F. (1987). Influential observations in survey sample. *Journal of Applied Statistics*, 14(2):143–152.