Open Journal of Statistics, **2012, 2, 281-296**
http://dx.doi.org/10.4236/ojs.2012.23034 Published Online July 2012 (http://www.SciRP.org/journal/ojs)

Subsampling Method for Robust Estimation of Regression Models

Min Tsao, Xiao Ling
Department of Mathematics and Statistics, University of Victoria, Victoria, Canada
Email: tsao@math.uvic.ca

Received March 29, 2012; revised April 30, 2012; accepted May 10, 2012

ABSTRACT

We propose a subsampling method for robust estimation of regression models which is built on classical methods such as the least squares method. It makes use of the non-robust nature of the underlying classical method to find a good sample from regression data contaminated with outliers, and then applies the classical method to the good sample to produce robust estimates of the regression model parameters. The subsampling method is a computational method rooted in the bootstrap methodology which trades analytical treatment for intensive computation; it finds the good sample through repeated fitting of the regression model to many random subsamples of the contaminated data instead of through an analytical treatment of the outliers. The subsampling method can be applied to all regression models for which non-robust classical methods are available. In the present paper, we focus on the basic formulation and robustness property of the subsampling method that are valid for all regression models. We also discuss variations of the method and apply it to three examples involving three different regression models.

Keywords: Subsampling Algorithm; Robust Regression; Outliers; Bootstrap; Goodness-of-Fit

1. Introduction

Robust estimation and inference for regression models is an important problem with a long history in robust statistics. Earlier work on this problem is discussed in [1] and [2]. The first book focusing on robust regression is [3] which gives a thorough coverage of robust regression methods developed prior to 1987. There have been many new developments in the last two decades. Reference [4] provides a good coverage on many recent robust regression methods. Although there are now different robust methods for various regression models, most existing methods involve a quantitative measure of the outlyingness of individual observations which is used to formulate robust estimators. That is, contributions from individual observations to the estimators are weighted depending on their degrees of outlyingness. This weighting by outlyingness is done either explicitly as in, for example, the GM-estimators of [5] or implicitly as in the MM-estimator of [6] through the use of ψ functions.

In this paper, we introduce an alternative method for robust regression which does not involve any explicit or implicit notion of outlyingness of individual observations. Our alternative method focuses instead on the *presence or absence of outliers* in a subset (subsample) of a sample, which does not require a quantitative characterisation of outlyingness of individual observations, and attempts to identify the subsample which is free of outliers. Our method makes use of standard non-robust classical regression methods for both identifying the outlier free subsamples and then estimating the regression model with the outlier free subsamples. Specifically, suppose we have a sample consisting of mostly "good data points" from an ideal regression model and some outliers which are not generated by the ideal model, and we wish to estimate the ideal model. The basic idea of our method is to consider subsamples taken without replacement from the contaminated sample and to identify, among possibly many subsamples, "good subsamples" which contain only good data points. Then estimate the ideal regression model using only the good subsamples through a simple classical method. The identification of good subsamples is accomplished through fitting the model to many subsamples with the classical method, and then using a criterion, typically a goodness-of-fit measure that is sensitive to the presence of outliers, to determine whether the subsamples contain outliers. We will refer to this method as the *subsampling method*. The subsampling method has three attractive aspects: 1) it is based on elements of classical methods, and as such it can be readily constructed to handle *all* regression models for which non-robust classical methods are available, 2) under certain conditions, it provides unbiased estimators for the *ideal* regression model parameters, and 3) it is

easy to implement as it does not involve the potentially difficult task of formulating the outlyningness of individual observations and their weighting.

Point (3) above is particularly interesting as evaluating the outlyingness of individual observations is traditionally at the heart of robust methods, yet in the regression context this task can be particularly difficult. To further illustrate this point, denote by $(\mathbf{X}_i, Y_i)$ an observation where $Y_i \in \mathbb{R}^1$ is the response and $\mathbf{X}_i \in \mathbb{R}^q$ the corresponding covariates vector. The outlyingness of obseravtion $(\mathbf{X}_i, Y_i)$ here is with respect to the underlying regression model, not with respect to a fixed point in $\mathbb{R}^{q+1}$ as is in the location problem. It may be an outlier due to the outlyingness in either $\mathbf{X}_i$ or Y_i or both. In simple regression models where the underlying models have nice geometric representations, such as the linear or multiple linear regression models, the outlyingness of an $(\mathbf{X}_i, Y_i)$ may be characterized by extending measures of outlyingness for the location problem through for example the residuals. But in cases where Y_i is discrete such as a binary, Poisson or multinomial response, the geometry of the underlying models are complicated and the outlyingness of $(\mathbf{X}_i, Y_i)$ may be difficult to formulate. With the subsampling methods, we avoid the need to formulate the outlyingness of individual observations but instead focus on the consequence of outliers, that is, they typically lead to a poor fit. We take advantage of this observation to remove the outliers and hence achieve robust estimation of regression models. It should be noted that traditionally the notion of an "outlier" is often associated with some underlying measure of outlyingness of individual observations. In the present paper, however, by "outliers" we simply mean data points that are not generated by the ideal model and will lead to a poor fit. Consequently, the removal of outliers is based on the quality of fit of subsamples, not a measure of outlyingness of individual points.

The rest of the paper is organized as follows. In Section 2, we set up notation and introduce the subsampling method. In Section 3, we discuss asymptotic and robustness properties of the subsampling estimator under general conditions not tied to a specific regression model. We then apply the subsampling methods to three examples involving three different regression models in Section 4. In Section 5, we discuss variations of the subsampling method which may improve the efficiency and reliability of the method. We conclude with a few remarks in Section 6. Proofs are given in the Appendix.

2. The Subsampling Method

To set up notation, let $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ be a realization of a random vector $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$ satisfying regression model

$$E(Y_i) = g(\mathbf{x}_i, \boldsymbol{\beta}), \tag{1}$$

where $Y_i \in \mathbb{R}^1$ is the response variable, $\mathbf{X}_i \in \mathbb{R}^q$ is the corresponding covariates vector, $g(\mathbf{x}_i, \boldsymbol{\beta}) : \mathbb{R}^q \to \mathbb{R}^1$ is the regression function and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the regression parameter vector. To accommodate different regression models, the distributions of Y_i and $\mathbf{X}_i$ are left unspecified here. They are also not needed in our subsequent discussions.

Denote by $S_N = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\}$ a contaminated sample of N observations containing n "good data" generated by model (1) and m "bad data" which are outliers not from the model. Here n and m are unknown integers that add up to N. Let S_n and S_m be the (unknown) partition of S_N such that S_n contains the n good data and S_m contains the m bad data. To achieve robust estimation of $\boldsymbol{\beta}$ with S_N, the subsampling method first constructs a sample S_g to estimate the unknown good data set S_n, and then applies a (non-robust) classical estimator to S_g. The resulting estimator for $\boldsymbol{\beta}$ will be referred to as the *subsampling estimator* or SUE for $\boldsymbol{\beta}$. Clearly, a reliable and efficient S_g which captures a high percentage of the good data points in S_n but none of the bad points in S_m is the key to the robustness and the efficiency of the SUE. The subsampling algorithm that we develop below is aimed at generating a reliable and efficient S_g.

2.1. The Subsampling Algorithm

Let A be a random sample of size n_s taken without replacement from S_N, which we will refer to as a *subsample* of S_N. The key idea of the subsampling method is to construct the estimator S_g for S_n by using a sequence of subsamples from S_N.

To fix the idea, for some $n_s \le n$, let $\{A_1, A_2, A_3, \cdots\}$ be an infinite sequence of independent random subsamples each of size n_s. Let $\{A_1^*, A_2^*, A_3^* \cdots\}$ be the subsequence of good subsamples, that is, subsamples which do not contain any outliers. Each of these sequences contains only a finite number of distinct subsamples. We choose to work with a repeating but infinite sequence instead of a finite sequence of distinct subsamples as that finite number may be very large and the infinite sequence set-up provides the most convenient theoretical framework as we will see below. Consider using the partial union

$$B_j = \bigcup_{i=1}^{j} A_i^* \tag{2}$$

to estimate the good data set S_n. Clearly, B_j is a subset of S_n. The following theorem gives the consistency of B_j as an estimator for S_n.

Theorem 1 *With probability one,* $\{A_1^*, A_2^*, A_3^*, \cdots\}$ *has infinitely many elements and*

$$P\left(B_\infty = S_n\right) = 1. \quad (3)$$

Proof: See the Appendix.

To measure the efficiency of B_j as an estimator for S_n, let W_j be the number of points in B_j. Since $B_j \subseteq S_n$, it is an efficient estimator of S_n if the ratio W_j/n is close to one. But for any finite j, W_j is a random variable. Hence we use the expectation of the ratio, which we denote by $E_F\left(B_j\right)$, to measure the efficiency of B_j. We have

$$E_F\left(B_j\right) = \frac{E\left(W_j\right)}{n}, \quad j = 1, 2, \cdots,$$

where $E\left(W_j\right)$ is the expected number of good data points picked up by B_j. The following theorem gives a simple expression of $E_F\left(B_j\right)$ in terms of n and n_s.

Theorem 2 *The efficiency of B_j in recovering good data points is*

$$E_F\left(B_j\right) = \frac{E\left(W_j\right)}{n} = 1 - \left(\frac{n - n_s}{n}\right)^j, \quad j = 1, 2, \cdots \quad (4)$$

Proof: See the Appendix.

Theorem 2 indicates that $E_F\left(B_j\right)$ converges to 1 as j goes to infinity. The convergence is very fast when $\left(n - n_s\right)/n$ is not close to 1, that is, the subsample size n_s is not too small relative to n. Hence for properly chosen n_s and j, B_j is an excellent estimator for S_n. Nevertheless, in practice while we can generate the sequence $\{A_1, A_2, A_3, \cdots\}$ easily, the subsequence of good subsamples $\left\{A_1^*, A_2^*, A_3^*, \cdots\right\}$ and hence B_j are not available as we do not have the means to determine whether a subsample A_i is a good subsample or a bad subsample containing one or more outliers. To deal with this, for a fixed $r^* \in \mathbb{N}$, the subsampling algorithm described below finds a sequence of r^* pseudo-good subsamples $\left\{A_{(1)}, A_{(2)}, \cdots, A_{(r^*)}\right\}$, and takes their union to form the estimator S_g for S_n.

Denote by Π a classical method for regression analysis, for example, the method of least squares. Denote by Γ an associated quantitative goodness-of-fit criterion, such as the mean squared error, AIC or BIC, which may be sensitive to the presence of outliers on a relative basis; that is, given two samples of the same size, one contains outliers and another does not, upon fitting regression model (1) with method Π to the two samples, Γ can be used to effectively identify the one that contains outliers. Denote by γ the numerical score given by the criterion Γ upon fitting the model (1) to a subsample A using method Π, and suppose that a small γ value means a good fit of model (1) to A. The subsampling algorithm finds pseudo-good subsamples $\left\{A_{(1)}, A_{(2)}, \cdots, A_{(r^*)}\right\}$ and forms S_g as follows.

ALGORITHM SAL (n_s, r^*, k)—Subsampling algorithm based on Π and Γ

For chosen (Π, Γ) and $n_s, r^*, k \in \mathbb{N}$ where $n_s \le n$:

Step 1: Randomly draw a subsample A_1 of size n_s from data set S_N.

Step 2: Using method Π, fit the regression model (1) to the subsample A_1 obtained in Step 1 and compute the corresponding goodness-of-fit score γ_1.

Step 3: Repeat Steps 1 and 2 for k times. Each time record $\left(A_j, \gamma_j\right)$, the subsample taken and the associated goodness-of-fit score at the *j*th repeat, for $j = 1, 2, \cdots, k$.

Step 4: Sort the k subsamples by their associated γ values; denote by $\gamma_{(1)}, \gamma_{(2)}, \ldots, \gamma_{(k)}$ the ordered values of γ_j where $\gamma_{(i)} \le \gamma_{(i+1)}$, and denote by $A_{(1)}, A_{(2)}, \ldots, A_{(k)}$ the correspondingly ordered subsamples. This puts the subsamples in the order of least likely to contain outliers to most likely to contain outliers according to the Γ criterion.

Step 5: Form S_g using the r^* subsamples with the smallest γ values, that is

$$S_g = \bigcup_{i=1}^{r^*} A_{(i)}. \quad (5)$$

We will refer to S_g as the *combined sample*. It estimates S_n using the pseudo-good subsamples $\left\{A_{(1)}, A_{(2)}, \cdots, A_{(r^*)}\right\}$ which is an approximate version of $\left\{A_1^*, A_2^*, \cdots, A_{r^*}^*\right\}$. Formally, we now define the subsampling estimator SUE for $\boldsymbol{\beta}$ as the estimator generated by applying method Π to the combined sample S_g. Although the method applied to S_g to generate the SUE does not have to be the same as the method used in the algorithm to generate S_g, for convenience we will use the same method in both places. The SUE does not have an analytic expression. It is an implicit function of n_s, r^*, k, Π, Γ and S_N, and as such we may write SUE ($n_s, r^*, k, \Pi, \Gamma, S_N$) when we need to highlight the ingredients.

It should be noted that while B_{r^*}, the union of $A_1^*, A_2^*, \cdots, A_{r^*}^*$, may be viewed as a random sample taken without replacement from S_n, the combined sample S_g may not be viewed as such even when it contains no outliers. This is because the r^* pseudo-good subsamples $\left\{A_{(1)}, A_{(2)}, \cdots, A_{(r^*)}\right\}$ are not independent due to the dependence in their γ-scores, which are the smallest r^* order statistics of the k γ-scores. Consequently, S_g is not a simple random sample. This complicates the distributional theory of the SUE but fortunately it has

little impact on its breakdown robustness as we will see later.

We conclude this subsection with an example typical of situations where we expect the algorithm to be successful in finding a good combined sample S_g for estimating S_n.

Example 1: Consider the following linear model

$$y = 3 + 5x + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2 = 4\right). \tag{6}$$

We generated $n = 18$ observations from model (6) and then added $m = 2$ outliers to form a data set of size $N = n + m = 20$. See **Figure 1(a)** for the scatter plot of this data set. To recover the 18 good data points, consider using the subsampling algorithm with the method of least squares as Π and the MSE $\left(SS_{\text{residual}} / \left(n_s - 2\right)\right.$ as the Γ criterion. For $n_s = 17$, there are 1140 such subsamples, 18 of which contain no outliers. We fitted the simple linear model using the method of least squares to all 1140 subsamples and computed the MSE for each fit. **Figure 1(b)** shows the plot of the 1140 MSEs versus the corresponding subsample labels. The MSEs of the 18 good subsamples are at the bottom, and they are substantially smaller than that of the other subsamples. If we are to run SAL $\left(n_s = 17, r^* = 2, k\right)$ for a suitably large k, the resulting combined sample S_g is likely the union of two good subsamples, recovering at least 17 of the 18 good data points.

For this example, the total number of distinct subsamples of size $n_s = 17$ is only 1140, which is not large.

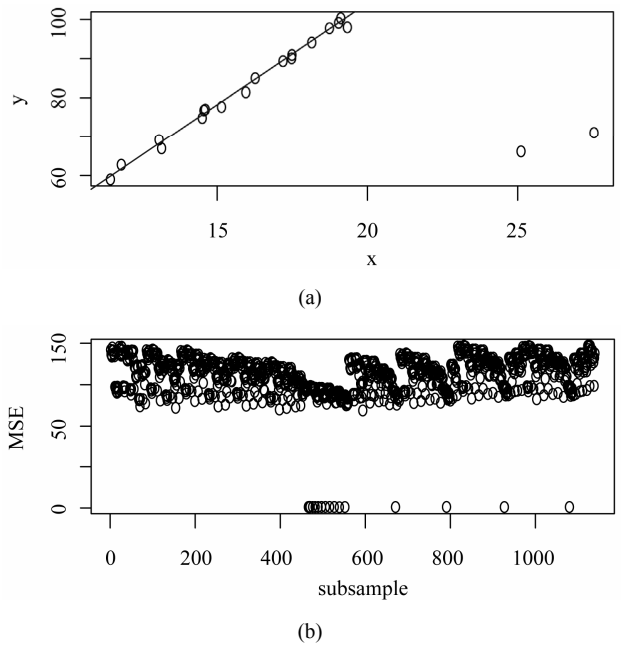

Figure 1. (a) Scatter plot of a contaminated sample of 20 from model (6) and the true regression line; (b) MSE from a least squares fit of a simple linear model to a subsample versus its label.

Instead of randomly generating subsamples in Steps 1-3 in SAL (n_s, r^*, k), we can modify it to going through all 1140 distinct subsamples. The combined sample S_g given by such a modified SAL $\left(n_s = 17, r^* = 2, k = 1140\right)$ will recover all 18 good data points in S_n. Consequently, the SUE reduces to the least squares estimator based on S_n and hence existing theory for the latter applies to the SUE.

2.2. Parameter Selection for the Subsampling Algorithm

In real applications, $\binom{N}{n_s}$ may be too large that going through all subsamples of size n_s is difficult. Thus we need to use random sampling in Steps 1-3 of SAL (n_s, r^*, k). The parameter selection discussed below is for SAL (n_s, r^*, k) with random sampling only. Also, the discussion below is independent of the underlying regression model. Thus we will not deal with the selection of Π and Γ (which is tied to the model) here but will address this for some regression models in Section 4 when we discuss specific applications.

The objective of SAL (n_s, r^*, k) is to generate a combined sample S_g to estimate the good data set S_n. To this end, we want to carefully select the parameter values to improve the chances that 1) the r^* pseudo-good subsamples $\left\{A_{(1)}, A_{(2)}, \cdots, A_{(r^*)}\right\}$ are indeed good subsamples from S_n and 2) S_g is efficient, that is, it captures a high percentage of points in S_n. The parameter selection strategy below centres around meeting these conditions.

1) Selecting n_s —The subsample size

Proper selection of n_s is crucial for meeting condition (1). The first rule for selecting n_s is that it must satisfy $m < n_s \le n$. Under this condition, $\left\{A_1, A_2, \cdots, A_k\right\}$ generated by Steps 1-3 are either good subsamples containing no outliers or subsamples containing a mix of good data points and outliers. The γ-score is the most effective in identifying good subsamples from such a sequence as they are the ones with small scores. If $n_s > n$, there will not be any good subsamples in the sequence. If $n_s < m$, there could be bad subsamples consisting of entirely outliers. When the outliers follow the same model as the good data but with a different $\boldsymbol{\beta}$, the γ-scores for bad subsamples of only outliers may also be very small. This could cause the subsampling algorithm to mistaken such bad subsamples as good subsamples, resulting in violations of condition (1).

In real applications, values of m and n may be unknown but one may have estimates of these which can be used to select n_s. In the absence of such estimates, we recommend a default value of "half plus one", *i.e.*

$n_s = \text{int}[0.5N]+1$ where int[x] is the integer part of x. This default choice is based on the assumption that $m < n$ which is standard in many robustness studies. Under this assumption and without further information, the only choice of n_s that is *guaranteed* to satisfy $m < n_s \le n$ is the default value.

Note that for each application, n_s must be above a certain minimum, *i.e.*, $n_s \ge p+1$. For example, if we let $n_s = 2$ when model (1) is a simple linear model where $p = 2$, we will get trivially perfect fit for all subsamples. In this case, the γ-score is a constant and hence cannot be used to differentiate good subsamples from the bad ones.

2) Selecting r^*—The number of subsamples to be combined in Step 5

The selection of r^* is tied to condition (2). For simplicity, here we ignore the difference between the pseudo-good subsamples $\left\{A_{(1)}, A_{(2)}, \cdots, A_{(r^*)}\right\}$ identified by the subsampling algorithm and the real good subsamples $\left\{A_1^*, A_2^*, \cdots, A_{r^*}^*\right\}$. This allows us to assume that S_g is the same as a B_{r^*} and use the efficiency measure of B_{r^*} for S_g. By (4),

$$E_F\left(S_g\right) = E_F\left(B_{r^*}\right) = 1-\left(\frac{n-n_s}{n}\right)^{r^*}. \tag{7}$$

For a given (desired) value of the efficiency $E_F\left(S_g\right)$, we find the r^* value needed to achieve this by solving (7) for r^*; in view of that r^* must be an integer, we have

$$r^* = \text{int}\left[\frac{\log\left(1-E_F\left(S_g\right)\right)}{\log\left(n-n_s\right)-\log\left(n\right)}\right]+1. \tag{8}$$

When n_s is the default value of $\text{int}[0.5N]+1$, the maximum r^* required to achieve a 99% efficiency ($E_F\left(S_g\right) = 99\%$) is 7. This maximum is reached when $n = N$.

In simulation studies, when r^* is chosen by (8), the observed efficiency, as measured by the actual number of points in S_g divided by n, tends to be lower than the expected value used in (8) to find r^*. This is likely the consequence of the dependence among the pseudo-good subsamples. We will further comment on this dependence in the next section.

3) Selecting k—The total number of subsamples to be generated

For a finite k, the sequence $\left\{A_1, A_2, \cdots, A_k\right\}$ generated by Steps 1-3 may or may not contain r^* good subsamples. But we can set k sufficiently large so that the probability of having at least r^* good subsamples, p^*, is high. We will use $p^* = 0.99$ as the default value for this probability. We now consider selecting k with given n_s, r^* and p^*.

Let p_g be the probability that A, a random subsample of size n_s from S_N, is a good subsample containing no outliers. Under the condition that $n_s \le n$, we have

$$p_g = \frac{\binom{n}{n_s}\binom{m}{0}}{\binom{n+m}{n_s}} > 0. \tag{9}$$

Now let T be the total number of good subsamples in $\left\{A_1, A_2, \cdots, A_k\right\}$. Then T is a binomial random variable with $T \sim Bin\left(k, p_g\right)$. Let $p\left(k,i\right) = P\left[T \ge i\right]$. Then

$$\begin{aligned} p\left(k,i\right) &= P\left(\text{at least } i \text{ good subsamples in } k\right) \\ &= 1-\sum_{j=0}^{i-1}\binom{k}{j}p_g^j\left(1-p_g\right)^{(k-j)}. \end{aligned} \tag{10}$$

Since the value of r^* has been determined in step [b] above, $p\left(k,r^*\right)$ is only a function of k and the desired k value is determined by p^* through

$$k = \arg\min_l\left\{p\left(l,r^*\right) \ge p^* = 0.99\right\}. \tag{11}$$

In real applications where m and hence n are unknown, we choose a *working proportion* $\alpha_0 \in (0.0.5)$, representing either our estimate of the proportion of outliers or the maximum proportion we will consider. Then use $m = \text{int}[N\alpha_0]$ to determine k. We will use 0.1 as the default value for α_0 which represents a moderate level of contamination.

Finally, k is a function of r^* and p_g, both of which are functions of n_s. Thus k is ultimately a function of only n_s. We may link the selection of all three parameters together by looking for the optimal n_s which will minimize k subject to the (estimated) constraint $m < n_s \le n$, where m and n are estimates based on α_0. Such an optimal n_s can be found by simply computing the k values corresponding to all n_s satisfying the constraint. We caution, however, that this optimal n_s may be smaller than $0.5N+1$. As such, it may not satisfy the real but unknown constraint of $m < n_s \le n$.

To give examples of parameter values determined through the above three steps and to prepare for the applications of the subsampling algorithm in subsequent examples, we include in **Table 1** the r^* values and k values required in order to achieve an efficiency of $E_F\left(S_g\right) = 99\%$ for various combinations of m and n values. To compute this table, both p^* and n_s are set to their default values of 0.99 and $0.5N+1$, respectively.

Table 1. The number of subsamples k and the number of good subsamples r^* required to achieve an efficiency of $E_F(S_g) = 99\%$. Exact values of m and n and default values of p^* and n_s were used for computing these k and r^* values.

Sample size N	Number of good data n	Number of outliers m	Subsample size n_s	r^*	Number of subsamples k
$N = 20$	$n = 20$	$m = 0$	$n_s = 11$	$r^* = 6$	$k = 6$
	$n = 18$	$m = 2$	$n_s = 11$	$r^* = 5$	$k = 58$
	$n = 16$	$m = 4$	$n_s = 11$	$r^* = 4$	$k = 383$
$N = 60$	$n = 60$	$m = 0$	$n_s = 31$	$r^* = 7$	$k = 7$
	$n = 54$	$m = 6$	$n_s = 31$	$r^* = 6$	$k = 1378$
	$n = 48$	$m = 12$	$n_s = 31$	$r^* = 5$	$k = 312{,}912$

The computationally intensive nature of SAL (n_s, r^*, k) can be seen in the case given by $(N, m, n) = (60, 12, 48)$, where to achieve an efficiency of 99% we need to fit the model in question to 312,912 subsamples of size 31 each. This could be a problem if it is time consuming to fit the model to each subsample. To visualize the relationship among various parameters of the algorithm, for the case of $(m, n) = (12, 48)$, **Figure 2** shows the theoretical efficiency of the algorithm $E_F(S_g)$ verses r^*. The dashed black line in the plot represents the 99% efficiency line. With subsample size $n_s = 31$, we see that the efficiency curve rises rapidly as r^* values increases. At $r^* = 5$, the efficiency reaches 99%. For this example, we also plotted the probability of having at least i good subsamples $p_i = p(k, i)$ verses i when the total number of subsamples is set at $k = 7000$ in **Figure 3(a)**. We see that at this k value, the probability of having at least $r^* = 5$ (required for 99% efficiency) is only about 0.96. To ensure a high probability of having at least 5 good subsamples, we need to increase k. In **Figure 3(b)**, we plotted the same curve but at $k = 312{,}912$ as was computed in **Table 1**. The probability of having at least 5 good subsamples is now at 0.99.

To summarize, to select the parameters n_s, r^* and k for SAL (n_s, r^*, k), we need the following information: 1) a desired efficiency $E_F(S_g)$ (default = 0.99), 2) a working proportion of outliers α_0 (default = 0.1) and 3) a probability p^* (default = 0.99) of having at least r^* good subsamples in k random subsamples of size n_s. We have developed an *R* program which computes the values of r^* and k for any combination of $(N, n_s, E_F(S_g), \alpha_0, p^*)$. The default value of this input vector is $(N, 0.5N + 1, 99\%, 0.1, 0.99)$. Note that the determination of the algorithm parameters does not depend on the actual model being fitted.

3. Asymptotic and Robustness Properties of the Subsampling Estimator

In this section, we discuss the asymptotic and robustness properties of the SUE under conditions not tied to a specific regression model.

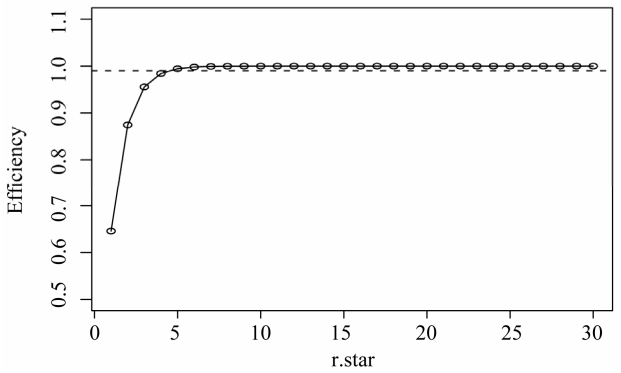

Figure 2. The efficiency of the subsampling algorithm SAL(n_s, r^*, k) as a function of the number of good subsamples r^* that form the combined sample S_g. The n_s is set at the default value. The dashed line is the 99% efficiency line.

3.1. The Asymptotic Distribution of the Subsampling Estimator

We first briefly discuss the asymptotic distribution of the SUE with respect to n_e, the size of the combined sample S_g. We will refer to n_e as the *effective sample size*. Although n_e is random, it is bounded between n_s and N. Under the default setting of $n_s = 0.5N + 1$, n_e will approach infinity if and only if N approaches infinity. Also, when the proportion of outliers and hence the ratio n/N are fixed, n_e will approach infinity if and only if n does. So asymptotics with respect to n_e is equivalent to that with respect to N or n.

Let $\hat{\boldsymbol{\beta}}$ be the SUE for $\boldsymbol{\beta}$ and consider its asymptotic distribution under the assumption that S_g may be viewed as random sample from the good data set S_n. Since S_n is a random sample from the ideal model (1), under the assumption S_g is also a random sample from model (1). Hence $\hat{\boldsymbol{\beta}}$ is simply the (classical) estimator

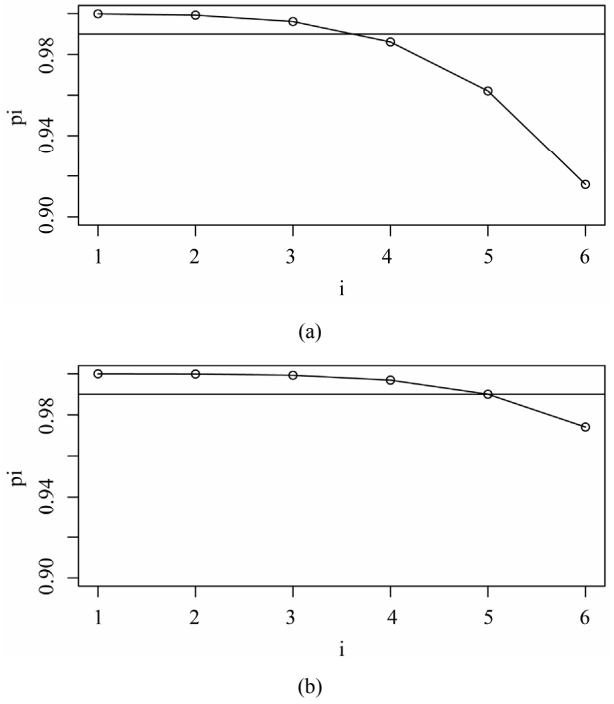

Figure 3. (a) The probability (pi) of having at least i good subsamples in k = 7000 subsamples; (b) The probability (pi) of having at least i good subsamples in k = 312,912 subsamples. The horizontal black line is the pi = 0.99 line.

given by method Π and the random sample S_g. Its asymptotic distribution is then given by the existing asymptotic theory for method Π. For example, for a linear model such as that in Example 1, the SUE $\hat{\boldsymbol{\beta}}$ generated by the least squares method will have the usual asymptotic normal distribution under the assumption. The asymptotic normal distribution can then be used to make inference about $\boldsymbol{\beta}$. Thus in this case, there is no need for new asymptotic theory for the SUE.

In some cases such as when it captures all the good data points, S_g may indeed be considered as a random sample from the good data. Nevertheless, as we have noted in Section 2.1 that in general, S_g is not a random sample due to a *selection bias* of the subsampling algorithm; the r^* subsamples forming S_g are the subsamples which fit the model the best according to the Γ criterion, and as such they tend to contain only those good data points which are close to the underlying regression curve. For the simple linear model, for example, good data points at the boundary of the good data band are more likely to be missed by S_g. Consequently, there may be less variation in S_g than that in a typical random sample from the model. This may lead to an underestimation of the model variance although the SUE for $\boldsymbol{\beta}$ may still be unbiased. The selection bias depends on the underlying model and the method Π. Its impact on the asymptotic distribution of the SUE $\hat{\boldsymbol{\beta}}$ needs to be studied on a case by case basis.

3.2. The Breakdown Robustness and the Breakdown Probability Function of the Subsampling Estimator

While a unified asymptotic theory for the SUE is elusive due to its dependence on the underlying model (1), the breakdown properties of the SUE presented below do not depend on the model and thus apply to SUEs for all models.

Consider an SUE($n_s, r^*, k, \Pi, \Gamma, S_N$) where n_s, r^*, k and N are fixed. Denote by α the proportion of outliers in S_N and hence $\alpha = m/N$. Due to the non-robust nature of the classical method Π, the SUE will break down if there is one or more outliers in S_g. Thus it *may* break down whenever α is not zero as there is a chance that S_g contains one or more outliers. It follows that the traditional notion of a breakdown point, the threshold α value below which an estimator will not break down, cannot be applied to measure the robustness of the SUE. The breakdown robustness of the SUE is better characterized by the *breakdown probability* as a function of α. This breakdown probability function, denoted by $BP(\alpha)$, is just the probability of *not* having at least r^* good subsamples in a random sequence of k subsamples of size n_s when the proportion of outliers is α. That is

$$BP(\alpha) = P\left(\text{fewer than } r^* \text{good subsamples in } k \middle| \alpha\right). \quad (12)$$

A trivial case arises when $n_s > n$. In this case, $BP(\alpha) = 1$ regardless of the α value. For the case of interest where $n_s \le n$, we have from (9)

$$p_g(\alpha) = \frac{\binom{n}{n_s}\binom{m}{0}}{\binom{n+m}{n_s}} > 0, \quad (13)$$

where $m = N\alpha$ and $n = N(1-\alpha)$. By (10), the breakdown probability function is

$$BP(\alpha) = \sum_{j=0}^{r^*-1} \binom{k}{j} p_g(\alpha)^j \left(1 - p_g(\alpha)\right)^{(k-j)}. \quad (14)$$

In deriving (14), we have assumed implicitly that the Γ criterion will correctly identify good subsamples without outliers. This is reasonable in the context of discussing the breakdown of the SUE as we can assume the outliers are arbitrarily large and hence any reasonable Γ criterion would be accurate in separating good subsamples from bad subsamples.

The concept of breakdown probability function can also be applied to traditional robust estimators. Let α^* be the breakdown point of some traditional robust esti-

mator. Then its breakdown probability function is the following step function

$$BP(\alpha)=\begin{cases}0, & \text{if } \alpha<\alpha^*,\\ 1, & \text{if } \alpha\geq\alpha^*.\end{cases} \quad (15)$$

Figure 4 contains the breakdown probability functions of three SUEs for the case of $N=60$. These SUEs are defined by the default values of $E(S_g)$, n_s and p^* with working proportions of outliers of $\alpha_0=0.1,0.2,0.4$, respectively. Their breakdown functions (in dotted lines) are identified by their working proportions. The breakdown function for the SUE with $\alpha_0=0.4$ is uniformly smaller than the other two and hence this SUE is the most robust. This is not surprising as it is designed to handle 40% outliers. For comparison, the breakdown probability function of a hypothetical traditional robust estimator (non-SUE) with a breakdown point of 0.3 is also plotted (in solid line) in **Figure 4**. Here we see that the SUE and the traditional robust estimator are complementary to each other; whereas the later will never breakdown so long as the proportion of outliers is less than its breakdown point but it will for sure breakdown otherwise, an SUE has a small probability of breaking down even when the proportion is lower than that it is designed to handle but this is compensated by the positive probability that it may not breakdown even when the proportion is higher. Incidentally, everything else being fixed, the k value associated with the SUE increases rapidly as the working proportion increases. The excellent robustness of the SUE for $\alpha_0=0.4$, for example, comes at a price of a huge amount of computation.

The breakdown probability function may be applied to select parameter k for the subsampling algorithm. Recall that in Section 2.2, after n_s and r^* are chosen and the working proportion of outliers α_0 is fixed, we find the value of k by using a predetermined p^*, the probability of having at least r^* good subsamples in k. In view of the breakdown probability function, this amounts to selecting k by requiring $BP(\alpha_0)=1-p^*$, which is a condition imposed on $BP(\alpha)$ at a single point α_0. An alternative way of selecting k is to impose a stronger condition on $BP(\alpha)$ over some interval of interest.

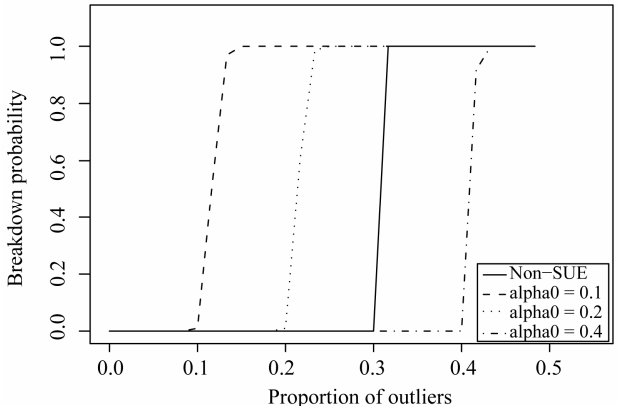

Figure 4. Breakdown probability functions (BPs) of three SUEs and one non-SUE for a data set of size N = 60. BPs for the SUE's designed to handle α_0 = 10%, 20% and 40% outliers are indicated by their associated α_0 value.

Note that everything else being equal, we can get a more robust SUE by using a larger k. For practical applications, however, we caution that a very large k will compromise both the computational efficiency of the subsampling algorithm and the efficiency of the combined sample S_g as an estimator of the good data set. The latter point is due to the fact that in practice, the subsamples forming S_g are not *independent* random samples from the good data set; in the extreme case where k goes to infinity, the subsample with the smallest γ-score will appear infinitely many times, and thus all r^* subsamples in the union of S_g are repeats of this same subsample. This leads to the lowest efficiency for S_g with $E_F(S_g)=n_s/N$. Thus when selecting the k value, it is necessary to balance the robustness and the efficiency of the SUE.

To conclude Section 3, we note that although the selection bias problem associated with the combined sample S_g can make the asymptotic theory of the SUE difficult, it has little impact on the breakdown robustness of the SUE. This is due to the fact that to study the breakdown of the SUE, we are only concerned with whether S_g contains any outliers. As such, the size of S_g and the possible dependence structure of points within are irrelevant. Whereas in Section 3.1 we had to make the strong assumption that S_g is a random sample from the good data in order to borrow the asymptotic theory from the classical method Π, here we do not need this assumption. Indeed, as we have pointed out after (14) that the breakdown results in this section are valid under only a weak assumption, that is, the Γ criterion employed is capable of separating good subsamples from subsamples containing one or more *arbitrarily large* outliers. Any reasonable Γ should be able to do so.

4. Applications of the Subsampling Method

In this section, we apply the subsampling method to three real examples through which we demonstrate its usefulness and discuss issues related to its implementation. For the last example, we also include a small simulation study on the finite sample behaviour of SUE.

An important issue concerning the implementation of the subsampling method which we have not considered in Section 2 is the selection of classical method Π and

goodness-of-fit criterion Γ. For linear regression and non-linear regression models, the least squares estimation (LSE) method and the mean squared error (MSE) are good choices for Π and Γ, respectively, as the LSE and MSE are very sensitive to outliers in the data. Outliers will lead to a poor fit by the LSE, resulting in a large MSE. Thus a small value of the MSE means a good fit. For logistic regression and Poisson regression models, the maximum likelihood estimation (MLE) method and the deviance (DV) can be used as Π and Γ, respectively. The MLE and DV are also sensitive to outliers. A good fit should have a small DV. If the ratio $\mathrm{DV}/(n_e - p)$ is much larger than 1, then it is not a good fit.

Another important issue is the proper selection of the working proportion of outliers or equivalently the (estimated) number of outliers m in the sample. This is needed to determine the r^* and k to run the subsampling algorithm. Ideally, the selected m value should be slightly above the true number of outliers as this will lead to a robust and efficient SUE. If we have some information about the proportion of outliers in the sample such as a tight upper bound, we can use this information to select m. In the absence of such information, we may use several values for m to compute the SUE and identify the most proper value for the data set in question. For m values above the true number of outliers, the SUE will give consistent estimates for the model parameters. Residual plots will also look consistent in terms of which points on the plots appear to be outliers. We now further illustrate these issues in the examples below.

Example 2: Linear model for stackloss data

The well-known stackloss data from Brownlee [7] has 21 observations on four variables concerning the operation of a plant for the oxidation of ammonia to nitric acid. The four variables are stackloss (y), air flow rate (x_1), cooling water inlet temperature (x_2) and acid concentration (x_3). We wish to fit a multiple linear regression model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

to this data. We use the LSE and MSE for Π and Γ, respectively, in the SUE. We also try three m values, $m = 2, 4$ and 6, which represent roughly 10%, 20% and 30% working proportion of outliers in the data. The subsample size is chosen to be the default size of $n_s = 11$. The corresponding values for r^* and k in the SAL and the estimates for regression parameters are given in **Table 2**. For comparison, **Table 2** also includes the estimates given by the LSE and MME, a robust estimator introduced in [6]. The residual versus fitted value plots for the LSE and SUE are in **Figure 5**. Since the regression parameter estimates for the SUE with $m = 4$ and

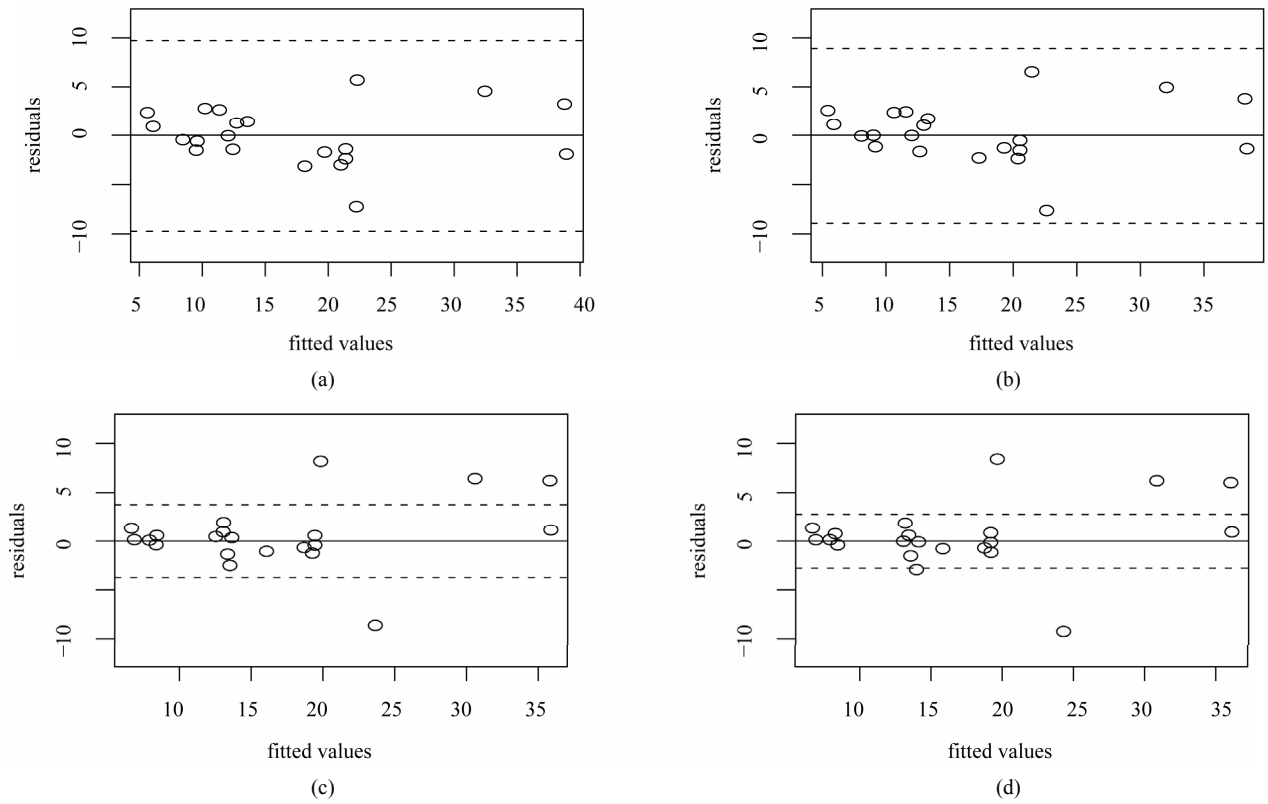

Figure 5. Residual versus fitted value plots for Example 2: (a) LSE; (b) SUE with *m* = 2; (c) SUE with *m* = 4; (d) SUE with *m* = 6. The dashed lines are $\pm 3\hat{\sigma}$, which are used to identify outliers.

Table 2. Regression parameter estimates with standard errors in brackets for Example 2.

parameter	LSE	MME	SUE ($m = 2$)	SUE ($m = 4$)	SUE ($m = 6$)
			$r^* = 6$	$r^* = 5$	$r^* = 4$
			$k = 57$	$k = 327$	$k = 2593$
β_0	−39.92 (11.90)	−41.52 (5.30)	−38.59 (10.93)	−37.65 (4.73)	−36.72 (3.65)
β_1	0.72 (0.13)	0.94 (0.12)	0.77 (0.13)	0.80 (0.07)	0.84 (0.05)
β_2	1.30 (0.37)	0.58 (0.26)	1.09 (0.35)	0.58 (0.17)	0.45 (0.13)
β_3	−0.15 (0.16)	−0.11 (0.07)	−0.16 (0.14)	−0.07 (0.06)	−0.08 (0.05)
σ	3.24	1.91	2.97	1.25	0.93
sample size	$n = 21$	$n = 21$	$n_e = 20$	$n_e = 17$	$n_e = 15$

$m = 6$ are consistent and the corresponding residual plots identify the same 4 outliers, $m = 4$ is the most reasonable choice. The effective sample size for S_g when $m = 4$ is $n_e = 17$, and hence this S_g includes all the good data points in the data and the SUE is the most efficient. It is clear from **Table 2** and **Figure 5** that the LSE and the SUE with $m = 2$ fail to identify any outliers and their estimates are influenced by the outliers. The robust MME identifies two outliers, and its estimates for β_1, β_2 and β_3 are slightly different from those given by the SUE with $m = 4$. Since the MME is usually biased in the estimation of the intercept β_0, the estimate of β_0 from the MME is quite different. This data set has been analysed by many statisticians, for example, Andrews [8], Rousseeuw and Leroy [3] and Montgomery *et al.* [9]. Most of these authors concluded that there are four outliers in the data (observations 1, 3, 4 and 21), which is consistent with the result of the SUE with $m = 4$.

Note that the SUE is based on the combined sample which is a trimmed sample. A large m value assumes more outliers and leads to heavier trimming and hence a smaller combined sample. This is seen from the SUE with $m = 6$ where the effective sample size n_e is 15 instead of 17 for $m = 4$. Consequently, the resulting estimate for the variance is lower than that for $m = 4$. However, the estimates for the regression parameters under $m = 4$ and $m = 6$ are comparable, reflecting the fact that under certain conditions the SUEs associated with different parameter settings of SAL algorithm are all unbiased.

Example 3: Logistic regression for coal miners data

Ashford [10] gives a data set concerning incidents of severe pneumoconiosis among 371 coal miners. The 371 miners are divided into 8 groups according to the years of exposure at the coal mine. The values of three variables, "years" of exposure (denoted by x) for each group, "total number" of miners in each group, and the number of "severe cases" of pneumoconiosis in each group, are given in the data set. The response variable of interest is the proportion of miners who have symptoms of severe pneumoconiosis (denoted by Y). The 8 group proportions of severe pneumoconiosis are plotted in **Figure 6(a)** with each circle representing one group. Since it is reasonable to assume that the corresponding number of severe cases for each group is a binomial random variable, on page 432 of [9] the authors considered a logistic regression model for Y, *i.e.*,

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

To apply subsampling method for logistic regression, we choose the MLE method and the deviance DV as Π and Γ, respectively. With $N = 8$ groups, we set $m = 1$ and 2 in the computation, and set the subsample size to $n_s = 5$. The corresponding values for r^* and k are $(r^*, k) = (4, 23)$ and $(r^*, k) = (3, 76)$ for $m = 1$ and 2, respectively. The original data set has no apparent outliers. In order to demonstrate the robustness of the SUE, we created one outlier group by changing the number of severe cases for the 27.5 years of exposure group from original 8 to 18. Consequently, the sample proportion of severe pneumoconiosis cases for this group has been changed from the initial 8/48 to 18/48. Outliers such as this can be caused, for example, by a typo in practice. The sample proportions with this one outlier are plotted in **Figure 6(b)**. The regression parameter estimates from various estimators are given in **Table 3**, where the M method is the robust estimator from [11]. The fitted lines given by the MLE, the SUE and the M method are also plotted in **Figure 6**. For both data sets, the SUE with $m = 1$ and 2 gives the same result. The SUE does not find any outliers for the original data set, while it correctly identifies the one outlier for the modified data set. For the original data set, the three methods give almost the same estimates for the parameters, and this is reflected by their fitted lines (of proportions) which are

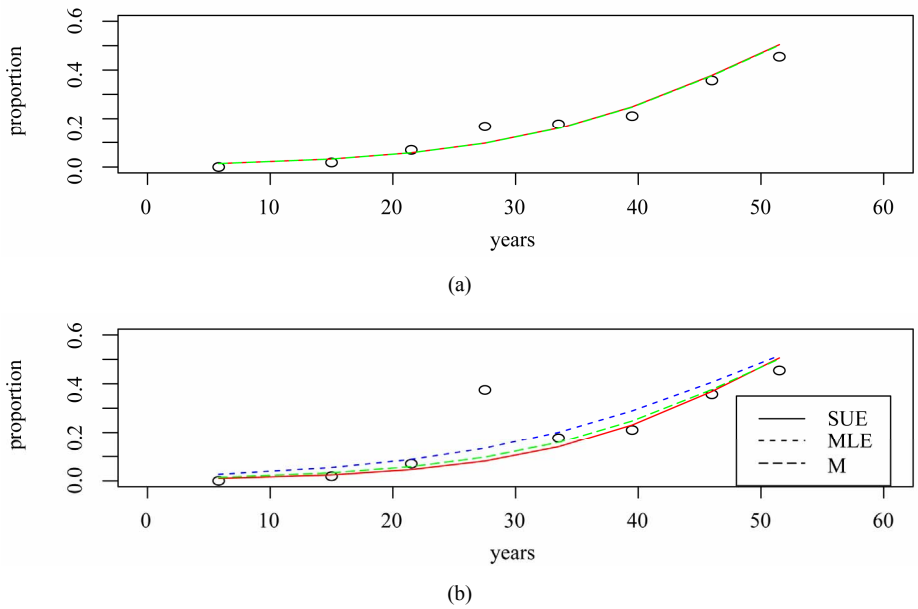

Figure 6. Sample proportions and fitted proportions for Example 3: (a) The original data set with no outlier group; (b) Modified data with one outlier group. Note that in (a) the fitted lines are all the same for the MLE, SUE and M methods, while in (b) they are different.

Table 3. Regression parameter estimates with standard errors in brackets for Example 3.

Parameter	Original data	With one Outlier group		SUE
	MLE	MLE	M	
β_0	−4.80 (0.57)	−4.07 (0.46)	−4.80 (0.59)	−5.24 (0.70)
β_1	0.09 (0.02)	0.08 (0.01)	0.09 (0.02)	0.10 (0.02)

nearly the same as can be seen in **Figure 6(a)**. For the modified data set, the SUE and the M method are robust, and their fitted lines are in **Figure 6(b)**. The outlier has little or no influence on these fitted lines.

Example 4: Non-linear model for enzymatic reaction data

To analyse an enzymatic reaction data set, one of the models that Bates and Watts [12] considered is the well-known Michaelis-Menton model, a non-linear model given by

$$y = \frac{\beta_0 x}{\beta_1 + x} + \varepsilon, \tag{16}$$

where β_0 and β_1 are regression parameters, and the error ε has variance σ^2. The data set has $N = 12$ observations of treated cases. Response variable y is the velocity of an enzymatic reaction and x is the substrate concentration in parts per million. To compute the SUE, we use the LSE method and the MSE for Π and Γ, respectively, and set $m = 2$ and $n_s = 7$ which lead to $r^* = 4$ and $k = 63$. **Figure 7(a)** shows the scatter plot of y versus x and the fitted lines for the LSE (dotted) and SUE (solid), and **Figure 7(b)** shows the residual plot from the SUE fit. Since there is only one mild outlier in the data, the estimates from the LSE and SUE are similar and they are reported in **Table 4**.

We also use model (16) to conduct a small simulation study to examine the finite sample distribution of the SUE. We generate 1000 samples of size $N = 12$ where, for each sample, 10 observations are generated from the model with $\beta_0 = 215$, $\beta_1 = 0.07$ and a normally distributed error with mean 0 and $\sigma = 8$. The other 2 observations are outliers generated from the same model but with a different error distribution; a normal error distribution with mean −30 and $\sigma = 1$. The two outliers are y outliers, and **Figure 7(c)** shows a typical sample with different fitted lines for the LSE and SUE. The estimated parameter values are also reported in **Table 4**. For each sample, the SUE estimates are computed with $n_s = 7$, $r^* = 4$ and $k = 63$. **Figure 8** shows the histograms for $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$ and the sample size of S_g. The dotted vertical lines are the true values for β_0, β_1, σ and $n = 10$. **Table 5** shows the biases and standard errors for the LSE and SUE based on the simulation study. The distributions of the SUE estimators $\hat{\beta}_0$ and

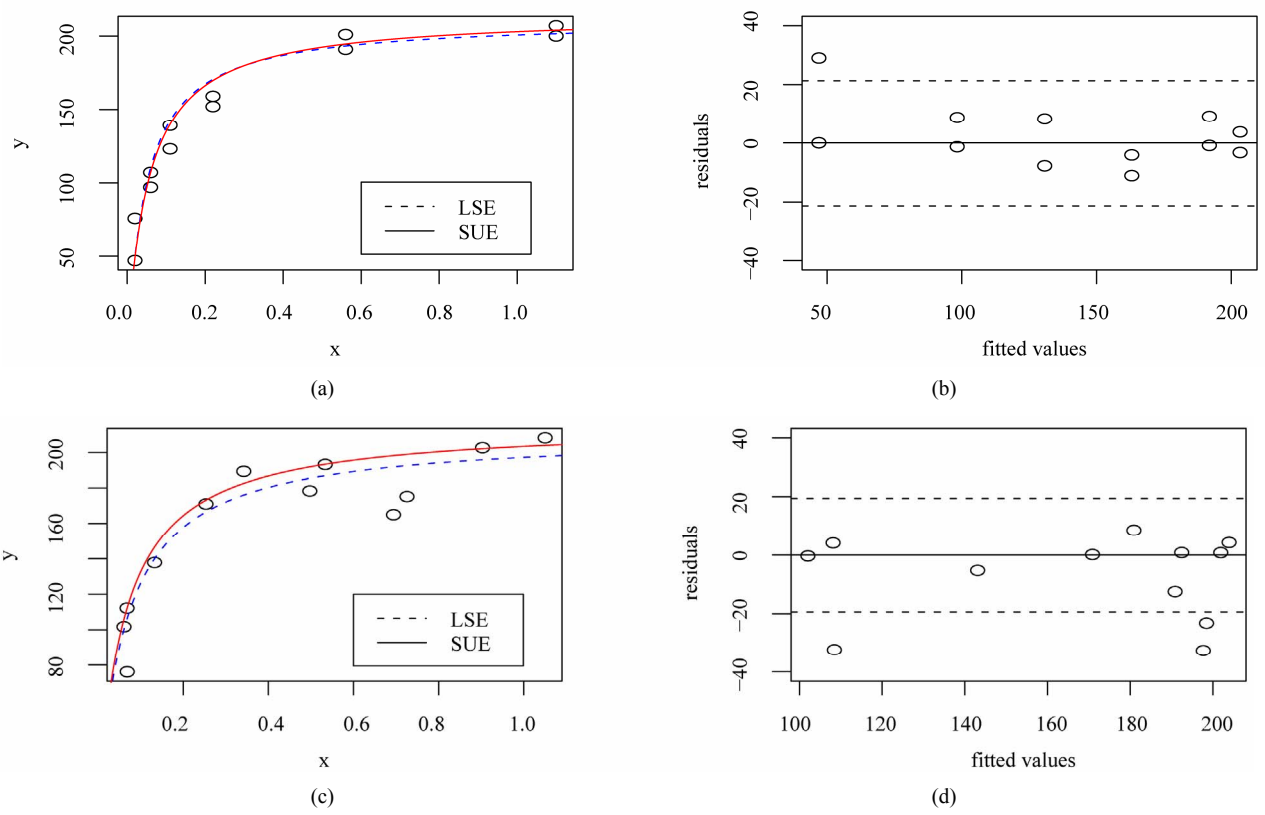

Figure 7. Fitted regression lines for Example 4: (a) The LSE and SUE lines for the real data set; (b) SUE fit residual plot for the real data set; (c) The LSE and SUE lines for the simulated data with outliers; (d) SUE fit residual plot for the simulated data. Dotted lines in plots (b) and (d) are the $\pm 3\hat{\sigma}$ lines.

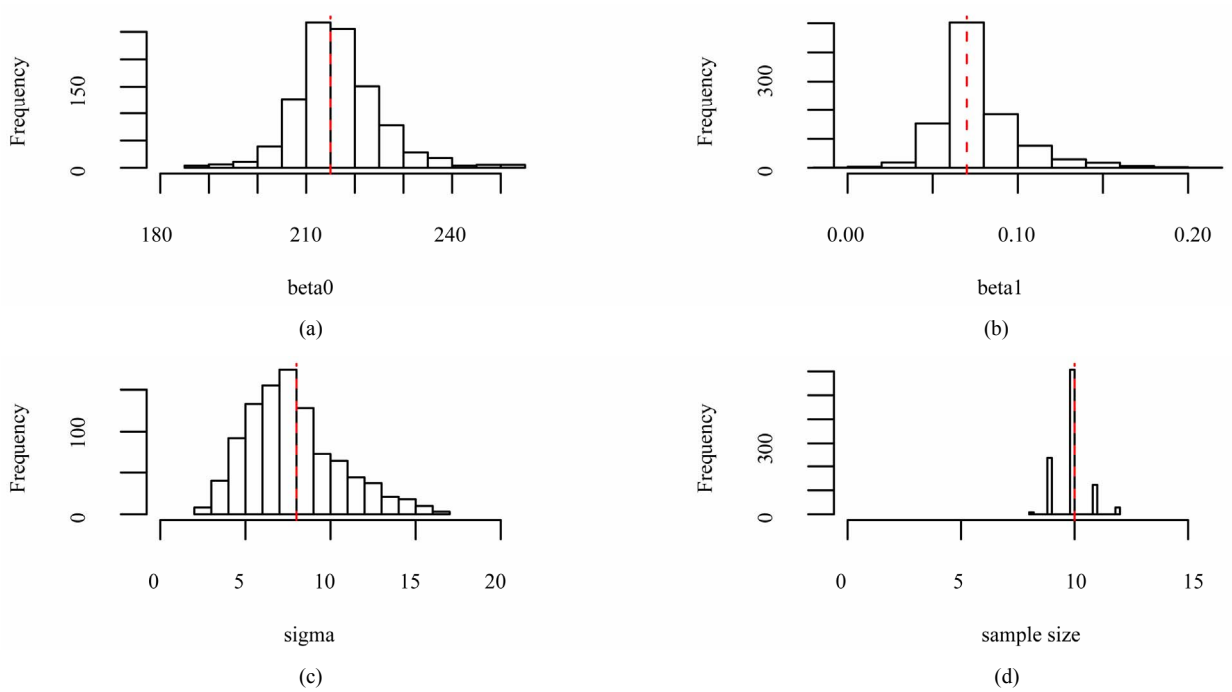

Figure 8. Histograms for the SUE: (a) Estimator $\hat{\beta}_0$; (b) Estimator $\hat{\beta}_1$; (c) Estimator $\hat{\sigma}$; (d) The sample size of S_g. The vertical dotted lines are the true values for β_0, β_1, σ and n.

Table 4. Regression estimates with the standard errors in brackets for Example 4.

Data set parameter	Original		Simulated	
	LSE (s.e.)	SUE (s.e.)	LSE (s.e.)	SUE (s.e.)
β_0	212.68 (6.95)	216.62 (4.79)	210.35 (9.03)	217.30 (4.55)
β_1	0.064 (0.008)	0.072 (0.006)	0.073 (0.014)	0.069 (0.007)
σ	10.93	7.10	15.23	6.47

Table 5. Results for the simulation study.

Parameter	LSE bias (s.e.)	SUE bias (s.e.)
β_0	−3.64 (8.84)	1.64 (8.93)
β_1	0.0059 (0.0238)	0.0052 (0.0266)
σ	5.70 (1.51)	−0.20 (2.71)

$\hat{\beta}_1$ are approximately normal, and the biases are much smaller than that of the LSE. That of the estimated variance also looks like a χ^2 distribution. The average effective sample size of S_g is 9.93, which is very close to the number of good data points $n=10$. There are a small number of cases where the effective sample size is 12. These are likely cases where the "outliers" generated are mild or benign outliers and are thus included in the combined sample.

5. Secondary Criterion and Other Variations of the Subsampling Algorithm

The 5-step subsampling algorithm SAL (n_s, r^*, k) introduced in Section 2 is the basic version which is straightforward to implement. In this section, we discuss modifications and variations which can improve its efficiency and reliability.

5.1. Alternative Stopping Criteria for Improving the Efficiency of the Combined Subsample S_g

In Step 5 of SAL (n_s, r^*, k), the first r^* subsamples in the sequence $A_{(1)}, A_{(2)}, \cdots, A_{(k)}$ are identified as r^* good subsamples and taken union of to form S_g. However, it is clear from the discussion on parameter selection in Section 2.2 that there are likely more than r^* good subsamples among the k generated by SAL (n_s, r^*, k). When there are more than r^* good subsamples, we want to use them all to form a larger and thus more efficient S_g. We now discuss two alternatives to the original Step 5 (referred to as Step 5a and Step 5b, respectively) that can take advantage of the additional good subsamples.

Step 5a: Suppose there is a cut-off point for the γ scores, say γ_C, such that the jth subsample is good if and only if $\gamma_{(j)} \le \gamma_C$. Then we define the combined subsample as

$$S_g = \bigcup_{\{j:\gamma_{(j)} \le \gamma_C\}} A_{(j)}. \tag{17}$$

Step 5b: Instead of a cut-off point, we can use $\gamma_{(1)}$ as a reference point and take the union of all subsamples whose γ scores are comparable to $\gamma_{(1)}$. That is, for a pre-determined constant $\kappa > 1$, we define the combined subsample as

$$S_g = \bigcup_{\{j:\gamma_{(j)} \le \kappa\gamma_{(1)}\}} A_{(j)}. \tag{18}$$

In both Steps 5a and 5b, the number of subsamples in the union are not fixed. The values of γ_C and κ depend on n, n_s, Π, Γ and the underlying model. Selection of γ_C and κ may be based on the distribution of γ-scores of good subsamples.

If either Step 5a or 5b is used instead of the original Step 5, we need to ensure the number of subsamples taken union of in (17) or (18) is no less than r^*. If it is less than r^*, then the criterion based on γ_C or κ may be too restrictive and it is not having the desired effect of improving the efficiency of the original Step 5. It may also be that the number of good subsamples is less than r^* and in this case, a re-run of SAL with a larger k is required.

Finally, noting that the r^* subsamples making up S_g in Step 5 may not be distinct, another way to increase efficiency is to use r^* distinct subsamples. The number of good subsamples in a sequence of k distinct subsamples follows a Hypergeometric (k, L_g, L_T) distribution, where L_g is the total number of good subsamples of size n_s and L_T the total number of subsamples of this size. Since L_T is usually much larger than L_g, the hypergeometric distribution is approximately a binomial distribution. Hence the required k for having r^* good subsamples with probability p^* is approximately the same as before.

5.2. Consistency of Subsamples and Secondary Criterion for Improved Reliability of the Subsampling Algorithm

Let $\boldsymbol{\beta}^i$ and $\boldsymbol{\beta}^j$ be the estimates given by (method Π

applied to) the *i*th and *j*th subsamples, respectively. Let $d\left(\boldsymbol{\beta}^i, \boldsymbol{\beta}^j\right)$ be a distance measure. We say that these two subsamples are *inconsistent* if $d\left(\boldsymbol{\beta}^i, \boldsymbol{\beta}^j\right) > d_c$ where d_c is a fixed positive constant. Conversely, we say that the two subsamples are consistent if $d\left(\boldsymbol{\beta}^i, \boldsymbol{\beta}^j\right) \le d_c$. Inconsistent subsamples may not be pooled into the combined sample S_g to estimate the unknown $\boldsymbol{\beta}$.

Step 5 of SAL (n_s, r^*, k) relies only on the γ-ordering of the subsamples to construct the combined sample S_g. In this and the two variations above, S_g is the union of (r^* or a random number of) subsamples with the smallest γ-scores. However, an S_g constructed in this manner may fail to be a union containing only good subsample as a small γ-score is not a sufficient condition for a good subsample. One case of bad subsamples with small γ-scores we have mentioned previously is that of a bad subsample consisting of entirely outliers that also follow the same model as the good data but with a different $\boldsymbol{\beta}$. In this case, the γ-score of this bad subsample may be very small. When it is pooled into the S_g, outliers will be retained and the resulting SUE will not be robust. Another case is when a bad subsample consists of some good data points and some outliers but the model happens to fit this bad subsample well, resulting in a very small γ-score. In both cases, the Γ criterion is unable to identify the bad subsamples but the estimated $\boldsymbol{\beta}$ based on these bad subsamples can be expected to be inconsistent with that given by a good subsample.

To guard against such failures of the Γ criterion, we use the consistency as a secondary criterion to increase the reliability of the SAL (n_s, r^*, k). Specifically, we pool only subsamples that are consistent through a modified Step 5, Step 5c, given below.

Step 5c: Denote by $\boldsymbol{\beta}^i$ the estimated value of $\boldsymbol{\beta}$ based on $A_{(i)}$ where $i = 1, 2, \cdots, k$ and let $d\left(\boldsymbol{\beta}^i, \boldsymbol{\beta}^j\right)$ be a distance measure between two estimated values. Take the union

$$S_g = \bigcup_{j=1}^{r^*} A_{(i_j)}, \tag{19}$$

where $A_{(i_j)}$ $(i_j = i_1, i_2, \cdots, i_{r^*})$ are the *first* r^* consistent subsamples satisfying

$$\max_{i_j \ne i_l} \left\{ d\left(\boldsymbol{\beta}^{i_j}, \boldsymbol{\beta}^{i_l}\right) \right\} \le d_c$$

for some predetermined constant d_c in $A_{(1)}, A_{(2)}, \cdots, A_{(k)}$.

The Γ criterion is the primary criterion of the subsampling algorithm as it provides the first ordering of the k subsamples. A secondary criterion divides the γ-ordered sequence into consistent subsequences and performs a grouping action (instead of an ordering action) on the subsamples. In principle, we can switch the roles of the primary and secondary criteria but this may substantially increase the computational difficulties of the algorithm. Additional criteria such as the range of the elements of $\boldsymbol{\beta}$ may also be added.

With the secondary criterion, the first r^* subsamples taken union of in Step 5c has an increased chance of all being good subsamples. While it is possible that they are actually *all* bad subsamples of the same kind (consistent bad subsamples with small γ-scores), this is improbable in most applications. Empirical experience suggests that for suitably chosen metric $d\left(\boldsymbol{\beta}^i, \boldsymbol{\beta}^j\right)$, threshold d_c and a reasonably large subsample size n_s (say $n_s = 0.5N + 1$), the first r^* subsamples are usually consistent, making the consistency check redundant. However, when the first r^* subsamples are not consistent and in particular when $i_{r^*} \gg r^*$, it is important to look for an explanation. Besides a poorly chosen metric or threshold, it is also possible that the data set actually comes from a mixture model. Apart from the original Step 5, a secondary criterion can also be incorporated into Step 5a or Step 5b.

Finally, there are other variations of the algorithm which may be computationally more efficient. One such variation whose efficiency is difficult to measure but is nevertheless worth mentioning here requires only $r^* = 1$ good subsample to start with. With $r^* = 1$, the total number of subsamples required (k) is substantially reduced, leading to less computation. Once identified, the one good subsample is used to test points outside of it. All additional good points identified through the test are then combined with the good subsample to form a combined sample. The efficiency of such a combined sample is difficult to measure and it depends on the test. But computationally, this approach is in general more efficient since the testing step is computationally inexpensive. This approach is equivalent to a partial depth function based approach proposed in [13].

6. Concluding Remarks

It is of interest to note the connections among our subsampling method, the bootstrap method and the method of trimming outliers. With the subsampling method, we substitute analytical treatment of the outliers (such as the use of the ρ functions in the M-estimator) with computing power through the elimination of outliers by repeated fitting of the model to the subsamples. From this point of view, our subsampling method and the bootstrap method share a common spirit of trading analytical treatment for intensive computing. Nevertheless, our subsampling method is not a bootstrap method as our objective is to identify and then make use of a single combined sample instead of making inference based on all bootstrap samples. The subsampling based elimination of outliers is also a generalization of the method of trimming outliers. Instead of relying on some measure of outlying-

ness of the data to decide which data points to trim, the subsampling method uses a model based trimming by identifying subsamples containing outliers through fitting the model to the subsamples. The outliers are in this case outliers with respect to the underlying regression model instead of some measure of central location and they are only implicitly identified as being part of the data points not in the final combined sample.

The subsampling method has two main advantages: it is easy to use and it can be used for any regression model to produce robust estimation for the underlying ideal model. There are, however, important theoretical issues that remain to be investigated. These include the characterization of the dependence structure among observations in the combined sample, the finite sample distribution and the asymptotic distribution of the SUE. Unfortunately, there does not seem to be a unified approach which can be used to tackle these issues for SUEs for all regression models; the dependence structure and the distributions of the SUE will depend on the underlying regression model as well as the method Π and criterion Γ chosen. This is yet another similarity to the bootstrap method in that although the basic ideas of such computation intensive methods have wide applications, there is no unified theory covering all applications and one needs to investigate the validity of the methods on a case by case basis. We have made some progress on the simple case of a linear model. We continue to examine these issues for this and other models, and hope to report our findings once they become available.

7. Acknowledgements

The authors would like to thank Dr. Julie Zhou for her generous help and support throughout the development of this work. This work was supported by a grant from the National Science and Engineering Research Council of Canada.

REFERENCES

[1] P. J. Huber, "Robust Statistics," Wiley, New York, 1981.

[2] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, "Robust Statistics: The Approach Based on Influence Functions," Wiley, New York, 1986.

[3] P. J. Rousseeuw and A. M. Leroy, "Robust Regression and Outlier Detection," Wiley, New York, 1987.

[4] R. A. Maronna, R. D. Martin and V. J. Yohai, "Robust Statistics: Theory and Methods," Wiley, New York, 2006.

[5] D. G. Simpson, D. Ruppert and R. J. Carroll, "On One-Step GM-Estimates and Stability of Inferences in Linear Regression," *Journal of the American Statistical Association*, Vol. 87, No. 418, 1992, pp. 439-450.

[6] V. J. Yohai, "High Breakdown-Point and High Efficiency Estimates for Regression," *Annals of Statistics*, Vol. 15, No. 2, 1987, pp. 642-656. doi:10.1214/aos/1176350366

[7] K. A. Brownlee, "Statistical Theory and Methodology in Science and Engineering," 2nd Edition, Wiley, New York, 1965.

[8] D. F. Andrews, "A Robust Method for Multiple Linear Regression," *Technometrics*, Vol. 16, No. 4, 1974, pp. 523-531.

[9] D. C. Montgomery, E. A. Peck and G. G. Vining, "Introduction to Linear Regression Analysis," 4th Edition, Wiley, New York, 2006.

[10] J. R. Ashford, "An Approach to the Analysis of Data for Semi-Quantal Responses in Biological Assay," *Biometrics*, Vol. 15, No. 4, 1959, pp. 573-581. doi:10.2307/2527655

[11] E. Cantoni and E. Ronchetti, "Robust Inference for Generalized Linear Models," *Journal of American Statistical Association*, Vol. 96, No. 455, 2001, pp. 1022-1030. doi:10.1198/016214501753209004

[12] D. M. Bates and D. G. Watts, "Nonlinear Regression Analysis and Its Applications," Wiley, New York, 1988.

[13] M. Tsao, "Partial Depth Functions for Multivariate Data," Manuscript in Preparation, 2012.

Appendix: Proofs of Theorems 1 and 2

Proof of Theorem 1:

Let p_g be the probability that random subsample of size n_s from S_N is a good subsample containing no outliers. Since $n_s \le n$, we have $p_g > 0$.

With probability 1, the subsequence $A_1^*, A_2^*, \cdots$ is an infinite sequence. This is so because the event that this subsequence contains only finitely many subsamples is equivalent to the event that there exists a finite l such that $A_l, A_{l+1}, \cdots$ contains no good subsamples. Since $p_g > 0$, the probability of this latter event and hence that of the former event are both zero. Further, under the condition that $A_j^* \subseteq S_n$, A_j^* may be viewed as a random sample of size n_s taken directly without replacement from S_n. Hence with probability 1, $A_1^*, A_2^*, \cdots$ is an infinite sequence of random samples from the finite population S_n. It follows that for any $z_j \in S_n$, the probability that it is in at least one of the A_i^* is 1. Hence $P(S_n \subseteq B_\infty) = 1$. This and the fact that $B_\infty \subseteq S_n$ imply the theorem.

Proof of Theorem 2:

We prove (4) by induction.

For $j = 1$, since $B_1 = A_1^*$ which contains exactly n_s points, W_1 is a constant and

$$E_F(B_1) = \frac{E(W_1)}{n} = \frac{n_s}{n}.$$

Hence (4) is true for $j = 1$.

To find $E_F(B_2)$, denote by $\bar{A}_1^*$ the complement of A_1^* with respect to S_n. Then $\bar{A}_1^*$ contains $n - n_s$ points. Since A_2^* is a random sample of size n_s taken without replacement from S_n, we may assume it contains U_1 data points from $\bar{A}_1^*$ and $n_s - U_1$ data points from A_1^*. It follows that U_1 has a hypergeometric distribution

$$U_1 \sim \text{Hyperg}(n, n - n_s, n_s), \tag{20}$$

with expected value

$$E(U_1) = n_s\left(\frac{n - n_s}{n}\right).$$

Since $B_2 = A_1^* \cup A_2^*$, its number of data points $W_2 = n_s + U_1$. Hence

$$E(W_2) = n_s + E(U_1) = n_s + n_s\left(\frac{n - n_s}{n}\right) = n\left[1 - \left(\frac{n - n_s}{n}\right)^2\right].$$

It follows that

$$E_F(B_2) = 1 - \left(\frac{n - n_s}{n}\right)^2,$$

and formula (4) is also true for $j = 2$.

Now assume (4) holds for some $(j-1) \ge 2$. We show that it also holds for j. Denote by $\#(A)$ the number of points in A. Then $W_j = \#(B_j) = \#(B_{j-1} \cup A_j^*) = W_{j-1} + U_{j-1}$, where U_{j-1} is the number of good data points in B_j but not in B_{j-1}. The distribution of U_{j-1} conditioning on $W_{j-1} = w_{j-1}$ is

$$U_{j-1} \big| (W_{j-1} = w_{j-1}) \sim \text{Hyperg}(n, n - w_{j-1}, n_s), \tag{21}$$

By (21) and the assumption that (4) holds for $j-1$, we have

$$\begin{aligned} E(W_j) &= E(W_{j-1}) + E\left(E\left(U_{j-1} \middle| W_{j-1}\right)\right) \\ &= E(W_{j-1}) + E\left[n_s\left(\frac{n - W_{j-1}}{n}\right)\right] \\ &= n_s + \frac{n - n_s}{n} E(W_{j-1}) \\ &= n_s + (n - n_s)\left[1 - \left(\frac{n - n_s}{n}\right)^{j-1}\right] \\ &= n - \frac{(n - n_s)^j}{n^{j-1}} = n\left[1 - \left(\frac{n - n_s}{n}\right)^j\right]. \end{aligned}$$

It follows that

$$E_F(B_j) = \frac{E(W_j)}{n} = 1 - \left(\frac{n - n_s}{n}\right)^j.$$

Thus (4) also holds for j, which proves Theorem 2.

