Elsevier Editorial System(tm) for
Computational Statistics and Data Analysis
Manuscript Draft

Manuscript Number: CSDA-D-16-01272R2

Title: Detection of influential points as a byproduct of resampling-based variable selection procedures

Article Type: Research Paper

Section/Category: II. Statistical Methodology for Data Analysis

Keywords: bootstrap; Grubbs' test; inclusion frequency; model averaging; outliers; subsampling.

Corresponding Author: Dr. Riccardo De Bin, Ph.D.

Corresponding Author's Institution: University of Oslo

First Author: Riccardo De Bin, Ph.D.

Order of Authors: Riccardo De Bin, Ph.D.; Anne-Laure Boulesteix, Professor; Willi Sauerbrei, Professsror

CSDA-D-16-01272R1
Title: Detection of influential points as a byproduct of resampling-based variable selection procedures
by R. De Bin, A. L. Boulesteix and W. Sauerbrei

Dear Editor,

we thank you for the opportunity to revise our paper. Please find here attached a new version of the manuscript that includes the additional changes suggested by a reviewer.

We look forward to hearing from you.

With best regards,

Riccardo De Bin
Anne-Laure Boulesteix
Willi Sauerbrei

RESPONSE TO REVIEWERS

We are grateful for the additional remarks of the Reviewer 1. This revised version of our manuscript contains the following changes to address the points raised by the reviewer:

- points 1 and 5 → we corrected the typos;
- points 2, 6, 7 and 8 → we rephrased the sentences as suggested by the reviewer;
- point 3 → we added a reference for the FSDA MATLAB toolbox as suggested by the reviewer;
- point 4 → we rewrote the whole sentence;
- point 9 → we removed the caption “observation 44” from Figure 7 as suggested by the reviewer.

Detection of influential points as a byproduct of resampling-based variable selection procedures

Riccardo De Bin

Department of Mathematics, University of Oslo
Department of Medical Informatics, Biometry and Epidemiology, University of Munich

Anne-Laure Boulesteix

Department of Medical Informatics, Biometry and Epidemiology, University of Munich

Willi Sauerbrei

Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg

Email addresses: debin@math.uio.no, postal address: Postboks 1053 Blindern 0316 Oslo (Norway), telephone:(+47)22855859 (Riccardo De Bin), boulesteix@ibe.med.uni-muenchen.de (Anne-Laure Boulesteix), wfs@imbi.uni-freiburg.de (Willi Sauerbrei)

Preprint submitted to Elsevier *July 3, 2017*

Detection of influential points as a byproduct of resampling-based variable selection procedures

Riccardo De Bin

Department of Mathematics, University of Oslo
Department of Medical Informatics, Biometry and Epidemiology, University of Munich

Anne-Laure Boulesteix

Department of Medical Informatics, Biometry and Epidemiology, University of Munich

Willi Sauerbrei

Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg

Abstract

Influential points can cause severe problems when deriving a multivariable regression model. A novel approach to check for such points is proposed, based on the variable inclusion matrix, a simple way to summarize results from resampling-based variable selection procedures. These procedures rely on the variable inclusion matrix, which reports whether a variable (column) is included in a regression model fitted on a pseudo-sample (row) generated from the original data (e.g., bootstrap sample or subsample). The variable inclusion matrix is used to study the variable selection stability, to derive weights for model averaged predictors and in others investigations. Concentrating on variable selection, it also allows understanding whether the presence of a specific observation has an influence on the selection of a variable. From the variable inclusion matrix, indeed, the inclusion frequency (I-frequency) of each variable can be computed only in the pseudo-samples (i.e., rows) which contain the specific observation. When the procedure is repeated for each

Email addresses: `debin@math.uio.no, postal address: Postboks 1053 Blindern 0316 Oslo (Norway), telephone:(+47)22855859` (Riccardo De Bin), `boulesteix@ibe.med.uni-muenchen.de` (Anne-Laure Boulesteix), `wfs@imbi.uni-freiburg.de` (Willi Sauerbrei)

observation, it is possible to check for influential points through the distribution of the I-frequencies, visualized in a boxplot, or through a Grubbs' test. Outlying values in the former case and significant results in the latter point to observations having an influence on the selection of a specific variable and therefore on the finally selected model. This novel approach is illustrated in two real data examples.

Keywords: bootstrap; Grubbs' test; inclusion frequency; model averaging; outliers; subsampling.

1. Introduction

In the construction of a statistical model, an important aspect to take into consideration is its stability. It is well known, indeed, that small perturbations in the data may lead to the selection of different models. For example, several papers show that variable selection procedures, such as backward elimination or forward selection, may provide very different sets of relevant variables, and consequently very different models, when applied to different bootstrap samples generated from the same dataset (Sauerbrei et al., 2015).

In the literature, different approaches have been proposed to handle this issue. From a variable point of view, resampling-based variable selection techniques can handle the instability issue by investigating the inclusion frequencies of the single variables (Gong, 1982; Chen & George, 1985). The idea is rather simple. Several pseudo-samples are generated via a resampling technique and a variable selection procedure is applied to select the best model in each of them. The proportion of models which contain the specific variable (inclusion frequency) is used as an indicator of the importance of the variable itself, and those variables with higher inclusion frequencies are used in the final model.

From a model point of view, model averaging is a technique which aims to deal with model uncertainty by fitting different models on the data and then summarizing their results. For example, in linear regression, a regression coefficient is estimated as a weighted mean of the corresponding estimates computed in each model. In particular, in the resampling-based approaches

Supplementary Material and the R-code to reproduce the results are available in a Web Appendix.

the weights are obtained by generating several pseudo-samples via a resampling technique and evaluating for how many of these pseudo-samples the different models are selected by a variable selection procedure. Other kinds of weights are based on information criteria, Mallows' criterion, etc. For a review on model averaging and on the different alternatives for the computation of the weights, we refer the reader to Wang et al. (2009). That paper, in particular, considers the frequentist approach. For a review about Bayesian model averaging, a classical reference is Hoeting et al. (1999).

Both resampling-based variable selection and resampling-based weights for model averaging require the application of a variable selection technique to several pseudo-samples. The goal of this paper is to show that the information collected in this part of the analysis can be used to check for influential points, such as outliers or single observations that have a high impact on the results. It is well known that influential points can cause problems when selecting a statistical model. For example, the inclusion or exclusion of a single or a few observations can have a dramatic effect on variables selected and on the issue of selecting linear or nonlinear function for a continuous variable (Royston & Sauerbrei, 2007). The literature on influential point detection is vast, and countless approaches have been proposed. For a simple and concise overview we refer the reader to Su & Tsai (2011) and references therein.

The detection of influential points as a byproduct of model-building procedures is not new. Tsao & Ling (2012), for example, exclude from the final model fitting procedure those observations that are not included in any of the pseudo-samples that lead to good models in terms of goodness-of-fit. A similar approach is used by Sauerbrei et al. (2015), who consider the selection probabilities of some "best models" and identify as influential points those observations which are able to modify these selection probabilities. Both approaches handle the influential point detection issue from a model point of view, ignoring the effect of these observations on the single variables. In this paper we consider the problem from a variable point of view, though maintaining a multivariable approach.

Finally, we mention Atkinson & Riani (2002), who also studied the effect of influential points from a model building point of view, using a forward search procedure (Atkinson & Riani, 2000, Ch. 2). We contrast our and their approaches in Section 4.1.5.

The paper is structured as follows. Section 2 presents two datasets later used as real examples. A brief introduction to model averaging and resampling-based variable selection is presented in Section 3, together with the description of our approach. The application of the method to the data is reported in Section 4. Finally, Section 5 contains a short discussion.

2. Data

2.1. Body fat data

The estimate of the percentage of body fat is considered a good indicator to assess the health of patients (see, e.g., Myint et al., 2014). Johnson (1996) presents a dataset in which the percentage of body fat (PBF) is collected from 252 men, together with the information about 13 further quantities, namely *age*, *weight*, *height* and 10 continuous body circumference measurements that are considered variables with potential influence on PBF. The data are publicly available at `http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book/Multivariable_Model-building/downloads/datasets/edu_bodyfat_both.zip`.

	BIC		$\alpha = 0.05$		AIC	
variable	in	out	in	out	in	out
age		✓		✓	✓	✓
weight	✓		✓		✓	
height		✓		✓		✓
neck					✓	✓
chest						✓
ab	✓	✓	✓	✓	✓	✓
hip					✓	
thigh					✓	
knee						
ankle						
biceps						
forearm	✓		✓		✓	✓
wrist	✓	✓	✓	✓	✓	✓

Table 1: Body fat data: result of a backward elimination procedure using three different selection criteria (BIC, significance level 0.05, AIC), with (in) and without (out) observation 39.

It is important to note that this dataset contains at least one influential point. Royston & Sauerbrei (2007), in particular, show that observation 39 highly influences the choice of the fractional polynomial function used to model the relationship between outcome and variables. Although some variables seem to have a non-linear effects on the outcome, we re-analyse this dataset under the assumption of linear effects. Non-linear effects are not that strong and this simplifying assumption seems acceptable for the main purpose of this paper.

To show the effect of observation 39 in a classical model-building procedure, we report in Table 1 the models obtained with backward elimination when this observation is included/excluded from the sample. Three common inclusion criteria are used. In this example, results are identical for BIC and $\alpha = 0.05$. As commonly seen in the literature (see, e.g., Sauerbrei et al., 2015), more variables are selected with AIC. We note that the presence/absence of observation 39 in the sample leads to substantially different models. The selections of variables *age*, *weight*, *height* and *forearm* are clearly affected.

2.2. Myeloma data

As an application of our method to a different kind of outcome, we also use a dataset with a time-to-event outcome. In particular, we consider a study on patients with multiple myeloma presented by Krall et al. (1975), in which the outcome is the survival time of the patients. The 16 variables are either binary or continuous. We consider the proportional hazard assumption acceptable, being this dataset analyzed several times in the literature by using the Cox model (see, e.g., Kuk, 1984; Chen & Wang, 1991). The sample size is small, consisting of 65 patients with 48 events. As for the body fat data, we use the simplifying assumption that the effect of continuous variables is linear. This dataset is also publicly available on the same website (`http://.../myeloma.zip`).

3. Methods

3.1. Resampling-based variable selection

One aim of a resampling-based variable selection is to select the relevant variables to include into a statistical model in a robust way, with the idea that the same model should be identified despite small perturbations in the data. In practice, a resampling technique, such as bootstrap or subsampling, is applied to the original dataset to generate several pseudo-samples, in order to mimic small perturbations in the data. As a sample with (bootstrap) or without (resampling) replacement from the original dataset, indeed, these pseudo-samples can be considered new instances of the data-generating mechanism, similar but not identical to the observed one. A variable selection technique, for example backward elimination, is then applied to each pseudo-sample. The proportion of pseudo-samples in which each variable is selected is called "inclusion frequency" and it is used to discriminate between relevant and irrelevant variables. The variables with higher inclusion frequencies are included in the final model, while the others are discarded. Table 2 reports an example of the computation of the inclusion frequencies. For further details and approaches to handle issues related to the dependence of inclusion frequencies among pairs of variables, see Sauerbrei & Schumacher (1992).

			variable					
pseudo-sample	V_1	V_2	V_3	$\dots$	V_{q-1}	V_q		model
1	1	0	1	$\dots$	0	1	$\rightarrow$	$\mathcal{M}_1$
2	0	1	1	$\dots$	0	0	$\rightarrow$	$\mathcal{M}_2$
3	1	0	1	$\dots$	0	1	$\rightarrow$	$\mathcal{M}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\rightarrow$	$\vdots$
B	1	0	1	$\dots$	0	0	$\rightarrow$	$\mathcal{M}_k$
inclusion frequency	0.961	0.243	1.000	$\dots$	0.000	0.693		

Table 2: Illustration of a variable inclusion matrix. It can be used to compute the resampling-based weights in a model averaging procedure (last column) or to compute the variable inclusion frequencies in a resampling-based variable selection procedure (last row).

3.2. Model averaging with resampling-based weights

The idea of model averaging consists in making inference on a parameter of interest by using several models instead of a single one. Consider K models $\mathcal{M}_1, \dots, \mathcal{M}_K$. The parameter estimate $\hat{\theta}$ is defined as the weighted average of the estimates computed across the K model ($\hat{\theta}_{\mathcal{M}_k}$), in formula

$$\hat{\theta} = \sum_{k=1}^{K} w_k \hat{\theta}_{\mathcal{M}_k}. \tag{1}$$

A highly relevant point is the choice of the weights w_k. In the literature several procedures have been proposed, for example based on information criteria (e.g. Buckland et al., 1997; Hjort & Claeskens, 2003) or Mallows' criterion (e.g. Hansen, 2007; Wan et al., 2010). Here we focus on weights based on a resampling approach, such as in, among others, Buckland et al. (1997); Augustin et al. (2005). As for resampling-based variable selection, a large number B of pseudo-samples are generated through a resampling technique and, to each pseudo-sample, a variable selection procedure is applied. In contrast to the previous approach, here the focus is not on the variables but on the resulting models. The proportion of time in which the model $\mathcal{M}_k$ is selected gives, for $k = 1, \ldots, K$, the weight w_k,

$$w_k = \frac{\#\mathcal{M}_k}{B}.$$

These weights are then used in formula (1). Although the inclusion matrix is the same as before (see Table 2), now the information is extracted on the direction of the rows (models).

Note that Hansen & Racine (2012) also used a resampling technique (in their case, jackknife) to derive the weights. Nevertheless, their approach relies on the estimate of the mean square error and therefore is theoretically different from the procedure described above.

3.3. Detection of possible influential points

3.3.1. From the inclusion matrix to the frequency matrix

We saw that both resampling-based variable selection and model averaging with resampling-based weights rely on an inclusion matrix. In each row, this matrix provides the information about which variables are included in the best model fitted on that particular pseudo-sample. For example, in a study with q variables, each row of the inclusion matrix is a q-dimensional vector containing 0 (variable not included) and 1 (variable included). The number of rows is arbitrary, and corresponds to the number of iterations performed. Table 2 reports an illustration of an inclusion matrix. As we saw above, in a resampling-based variable selection procedure this matrix is used to compute the inclusion frequencies for the variables (column averages), in a model averaging procedure to compute the weights (each row corresponds to a model).

Since each row corresponds to a pseudo-sample, the inclusion matrix also provides us with important information about the relationship between variables and observations. In addition to the inclusion/exclusion of the variables in the selected model, indeed, for each row we know which observations belong to the particular pseudo-sample and which do not. Combining these two aspects, we can evaluate the effect of a specific observation on the inclusion frequencies of the variables. For each observation i, we can estimate inclusion frequencies of all variables separately for samples including or excluding i. For a variable V_j, the two frequencies should be similar if the observation i has no effect on its inclusion and different if i has an influence on the inclusion of V_j.

Let us focus on the inclusion frequencies obtained by considering only the pseudo-samples in which a specific observation is included. For each observation $i = 1, \dots, n$, we compute these inclusion frequencies (hereafter, "I-frequencies", where "I" stands for "in") for all variables, obtaining a q-dimensional vector in which each entry corresponds to one variable (q is the number of variables). By merging these vectors, we obtain a $n \times q$ matrix of I-frequencies (hereafter, "I-frequency matrix"), as that reported in Table 3. In this example, in the pseudo-samples in which observation x_1 is included (first row), the variable V_1 is selected 0.969 of the times, V_2 0.015, and so on.

observation			variable			
included	V_1	V_2	V_3	$\dots$	V_{q-1}	V_q
1	0.969	0.015	0.553	$\dots$	0.000	0.292
2	1.000	0.030	0.492	$\dots$	0.000	0.376
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$n-1$	1.000	0.015	0.603	$\dots$	0.000	0.361
n	0.984	0.092	0.569	$\dots$	0.000	0.276

Table 3: Illustration of a I-frequency matrix. For each variable (column), it reports its I-frequencies, i.e. the inclusion frequency computed only on pseudo-samples in which a specific observation (row) is included.

3.3.2. I-frequency matrix and detection of influential points

If there is no influential point in the sample, we expect the values in the column of the I-frequency matrix to be very similar to each other. Conversely, the effect of an influential point would be visible in values that are strongly separated from the rest. Let us consider, as an example, an influential point, let say x_i, which strongly influences the significance of a variable V_j, in the sense that it forces V_j to enter into the model. Focusing on the column related to V_j, we would expect in the i-th row of the I-frequency-matrix a value much larger than all other values present in the same column.

Visualization. The easiest way to identify possible influential points is to plot the column values of the I-frequency matrix in boxplots, and take advantage of what Friedman and Tukey call "the human gift for pattern recognition" (Friedman & Tukey, 1974). The boxplot is a simple and effective tool to display the I-frequencies of a variable and to identify those that are far from the median value. In particular, in the standard way of drawing a boxplot, the extreme observations are not included in the whiskers and are plotted as separated points. Usually, this is done for points farther than 1.5 times the interquartile range from the first/third quartile. The farthest points are the values we are interested in, because they represent the most anomalous inclusion frequencies. One can then easily go back to the frequency matrix and identify the rows which correspond to these values, and, consequently, which are the possible influential points. In the case of no influential points, instead, we would expect no strongly separated points, i.e. a plot in which all values would be included or would be close to the boxplot's whiskers. Note, however, that identifying possible outliers among the points outside the whiskers is a delicate task, and more objective criteria may be necessary (see also Section 3.3.3).

Remark. The column variance of the I-frequency matrix can also be seen as an indicator of the "trustworthiness" of the variable inclusion frequency. Smaller variance, indeed, means an inclusion frequency that does not change too much in the case of small perturbations in the data. If for any reason we are in doubt whether a variable should or should not be included in the model, the variance may be a further argument to support our choice. For example, in the case of two correlated variables with similar inclusion frequencies, we may prefer to select that for which we obtain a smaller variance, because less influenced by small perturbations in the data.

3.3.3. Grubbs' tests

Although several researchers advocate graphical investigations to detect influential points, in some extends it may be advantageous to rely on a statistical test. From our point of view, we need to test whether the most extreme (i.e., farthest from the median frequency) I-frequency is an outlier for each variable. In the case of a positive answer, it would mean that one single observation, let us say $x_{(n)}$, is able to change the inclusion or exclusion of a variable in the model in a significant way. In other words, that $x_{(n)}$ may be an influential point. In order to evaluate the influence of each observation on each variable, we analyze the I-frequency matrix column by column. In this way, we can simply apply to each column a simple univariate test, such as the Dixon's Q (Dixon, 1950) and the Grubbs' G (Grubbs, 1950). Due to the dependence of the former to the sample size, here we use the latter. It is worth stressing, in any case, that our analysis is meant as explorative. Once the aforementioned $x_{(n)}$ has been selected by our procedure, it is the responsibility of the practitioner to evaluate the exact nature of the observation (i.e., whether it is actually an influential point).

Given a sample $x_1, \ldots, x_n$ from a Gaussian distribution, the Grubbs' test rejects the null hypothesis, defined as the absence of outliers, if

$$\max_{i=1,\ldots,n} \frac{|x_i - \bar{x}|}{s} > C(\alpha, n) = (n-1)\sqrt{\frac{t^2_{1-\alpha/(2n),n-2}}{(n-2+t^2_{\alpha/(2n),n-2})}},$$

where $\bar{x}$ denotes the sample mean, s the estimated standard deviation and $t_{1-\alpha/(2n),n-2}$ the quantile $1-\alpha/(2n)$ of a t distribution with $n-2$ degrees of freedom. Here α is the significance level on which the test is conducted; since we repeat the test for each variable, it may be necessary to implement a correction for the multiplicity of the tests.

Visualization. For an easy identification of the influential points, it may be convenient to visualize the results in a graphic. Our suggestion is to plot, for each variable (i.e., for each column of the I-frequency matrix), the standardized I-frequency. This value is strictly related to the test statistic of the Grubbs' test, with the difference that we do not consider the absolute value but simply the difference between the value and the mean. If one value is outside the bands $\pm C(\alpha, n)$ it means that the I-frequency is an outlier and the corresponding observation may be an influential point. Please note that the Grubbs' test is constructed to identify the presence of one outlier. In general, a new critical value $C(\alpha, n)$ should be considered in the case of multiple outliers, namely

$$C(\alpha, n, k) = (n-k)\sqrt{\frac{t^2_{1-\alpha/(2(n-k+1)),n-k-1}}{(n-k-1+t^2_{\alpha/(2(n-k+1)),n-k-1})}},$$

where k indicates the number of outliers whose presence in the sample one wants to test. Nevertheless, for reasonably large sample size ($n > 50$), the critical value does not change much with k and the original $C(\alpha, n)$ can be used.

Remark. Note that the I-frequencies do not follow a Gaussian distribution, which is an assumption of the Grubbs' test. Their distribution may be better described by a beta distribution with accumulation points on the boundaries (0 and 1). Nevertheless, the beta distribution can be approximated by a Gaussian distribution when its coefficients are sufficiently large, i.e., when the data points are far from the boundaries. In fact, we are only interested in these cases. I-frequencies close to 0, indeed, are related to irrelevant variables, which should not be included into the final model. On the other extreme, I-frequencies close to 1 are typical of strong variables, which are almost always included in the model. In these two cases, the possible presence of an influential point would not change our decision to include or exclude the variable from the final model. In contrast, the dependence among the I-frequencies, which are computed on the same pseudo-samples, is not a problem. It has been shown that Grubbs' test is robust against deviation from independence (Srivastava, 1980).

3.4. Effect of the choice of the resampling technique

The construction of the inclusion matrix needs the implementation of a resampling technique to generate the pseudo-samples. Historically, bootstrap (Efron, 1979) has been the most used approach. It generates the pseudo-samples by sampling with replacement n observations (i.e. the original sample size) from the original sample. Alternatives such as subsampling (Hartigan, 1969) have been also considered (see, e.g. Meinshausen & Bühlmann, 2010). Subsampling consists of sampling without replacement $m < n$ observations from the original sample.

The choice of the resampling technique should be driven by considerations on the model-building procedure, both for model averaging and for resampling-based variable selection (for a recent study in the latter case, see De Bin et al., 2016). For example, when there are variables with different numbers of categories, the use of the bootstrap may cause misleading results (Rospleszcz et al., 2016). However, from an influential point detection point-of-view, the bootstrap gives the possibility to separately consider pseudo-samples by the number of times (e.g.. 0, 1, more than 1) they include an observation i (Royston & Sauerbrei, 2008, Section 8.5.1). For this reason, we consider the bootstrap in our study.

3.5. Software

The following analysis have been computed using R (R Core Team, 2016) and the FSDA MATLAB toolbox (Riani et al., 2012). The R-code implementing the analyses is reported in the Web Appendix and the data are publicly available. That allows reproducibility of our study. Moreover, we plan to upload soon a specific R-package to the CRAN.

4. Results

4.1. Body fat data

From the original body fat sample, we generate 2000 bootstrap samples. To these pseudo-samples we applied a backward elimination procedure with significance level $\alpha = 0.05$. As a result, we obtain a 2000×13 inclusion matrix. We explained that this matrix can be used to perform variable selection or to compute the weights for a model averaging procedure. Here, instead, we use it to generate the I-frequency matrix (as Table 3) and to check the possible presence of influential points.

4.1.1. I-frequency matrix

The I-frequency matrix is a 252×13 matrix whose columns report the I-frequencies for the 13 variables and whose rows correspond to the observations as explained above.

Figure 1 shows boxplots of the inclusion frequencies (see Section 3.3.1). We note some points that are far from the respective median frequencies and, in general, from all other I-frequencies. This fact is a sign of the possible presence of influential points in the data. In particular, for the variables *weight*, *height*, *chest* and *forearm* we note four points (one for each variable) with this characteristic. All four points correspond to the inclusion frequencies computed on pseudo-samples which include observation 39. As stated in Section 2.1, observation 39 is a known outlier, and our method is able to correctly identify it. Interestingly, there seems to be an observation (observation 221) which also has an effect on the inclusion of *weight*, in an opposite direction than observation 39.

Figure 1 contains further information. First of all, we can see the effect of the influential point in the model building process. It is clear that observation 39 leads to an overestimation of the importance of *weight* and *forearm*, and an underestimation of *height* and *chest*. Secondly, we can visualize the strengths of the variables. We note, for example, that *ab* (abdominal circumference) is a

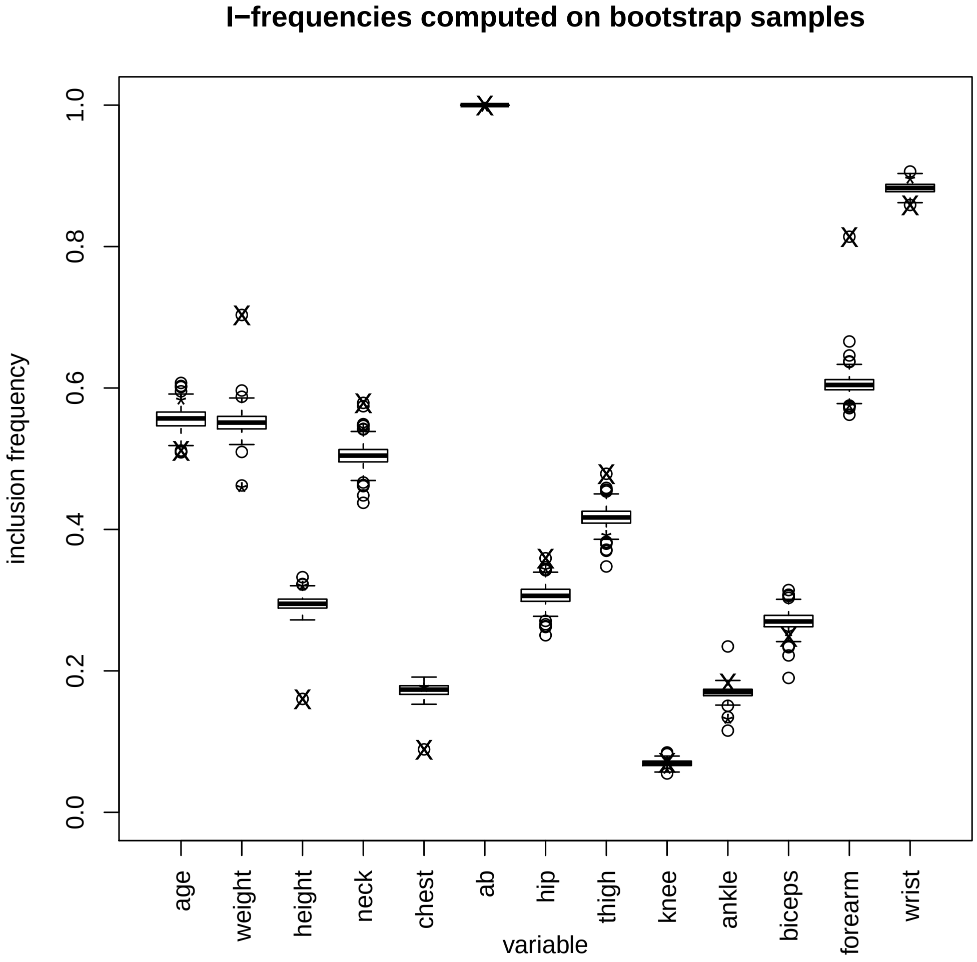


Figure 1: Body fat data: boxplots that summarize the I-frequencies obtained for the 13 variables. The symbols "X" and "*" denote the I-frequencies obtained in pseudo-samples (here generated by using bootstrap) which include observation 39 and 221, respectively.

very strong variable, which is always included in the model, while *knee*, which is included in only a few models (low inclusion frequency), seems irrelevant. Moreover, this figure allows us to compare the variances of the inclusion frequencies. As mentioned before (Section 3.3.2), this information may be useful in the model building procedure.

4.1.2. Grubbs' tests

As described in Section 3.3.3, we can perform univariate tests on the I-frequencies for each variable. To visualize the results of the tests, we show boxplots of the standardized I-frequencies (observed test statistics) and draw the rejection region for the Grubbs' test with a significance level of 99% (Figure 2).

It is fairly simple to detect possible outliers. Note that the observed values for the test statistics for observation 39 are very improbable under the null hypothesis (no outlier), i.e., the points are deeply inside the rejection region and very far from its boundaries. This is true in particular when considering the variables *weight*, *height* and *forearm*. Moreover, observation 221 shows potential to be an influential point, with effect on the selection of *weight*. Although the corresponding point is not so distant from the rejection region boundary, we should take into consideration that its position is influenced by that of observation 39, which has opposite effect. Some pseudo-samples which include observation 221 also contain observation 39, and tend to increase the value of this point. Despite the masking effect of observation 39, observation 221 is inside the rejection region.

Further points are inside the rejection region, but they are related to the variables *ankle* and *biceps*, which have low inclusion frequencies (see Figure 1). Therefore, the effect of the observations related to these points are not really interesting from a model building point of view, as *ankle* and *biceps* are not included in any case into the final model. Note that the lack of information on the strengths of the variables is a drawback of a plot based on standardized I-frequencies.

4.1.3. O-frequencies and multiple presences

As stated in Section 3.3.1, from the inclusion matrix we can compute, for each observation, the inclusion frequencies based on pseudo-samples which contain the specific observation (I-frequencies) or based on those which do not contain it. The choice of using the former was arbitrary, and potentially one can prefer the use of the latter (let us call them O-frequencies, where "O" stands for "out"). We report in the Web Appendix (Figures A.1 and A.2) the same graphics of Figures 1 and 2 when the O-frequencies are used instead of the I-frequencies. We note that, in this specific example, we obtain similar results. By using the O-frequencies, one should only remember that a smaller O-frequency for an observation x_i means an increasing effect on the inclusion of a variable. Graphically, a lower point means an higher inclusion

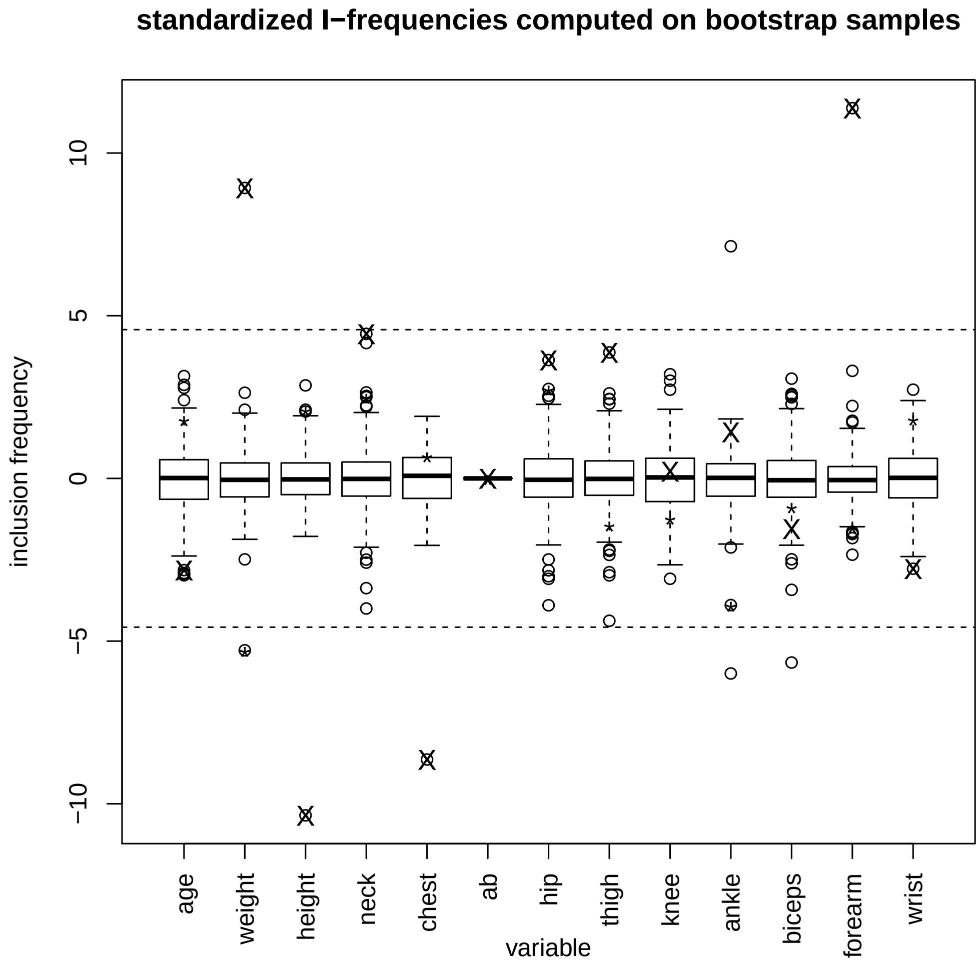


Figure 2: Body fat data: boxplots that summarize the standardized I-frequencies (computed on bootstrap pseudo-samples) obtained for the 13 variables. The symbols "X" and "*" denote the I-frequencies obtained in pseudo-samples (here generated by using bootstrap) which include observation 39 and 221, respectively. The dashed lines delimit the 99% rejection region of a Grubbs' test (including a correction for the multiplicity of the tests).

frequency. This seems counterintuitive and may generate confusion. For this reason, we preferred to use I-frequencies.

The O-frequencies, however, may prove useful in a different analysis. As stated in Section 3.4, when we use bootstrap as a resampling technique, we can separate the pseudo-samples in which a specific observation is included only one time from those in which it is included two or more times. We can then compute two separate inclusion frequencies, that we call I-frequencies-1 and I-frequencies-M, respectively. Together with the O-frequencies (i.e., frequencies computed on bootstrap samples in which the specific observation is included 0 times), these inclusion frequencies can provide us with additional information on the effect of the observation on the inclusion or exclusion of a variable from the statistical model. Pseudo-samples containing two or more times the specific observation are grouped together due to the relatively small amount of cases in which a single observation is repeated three or more times in a bootstrap sample.

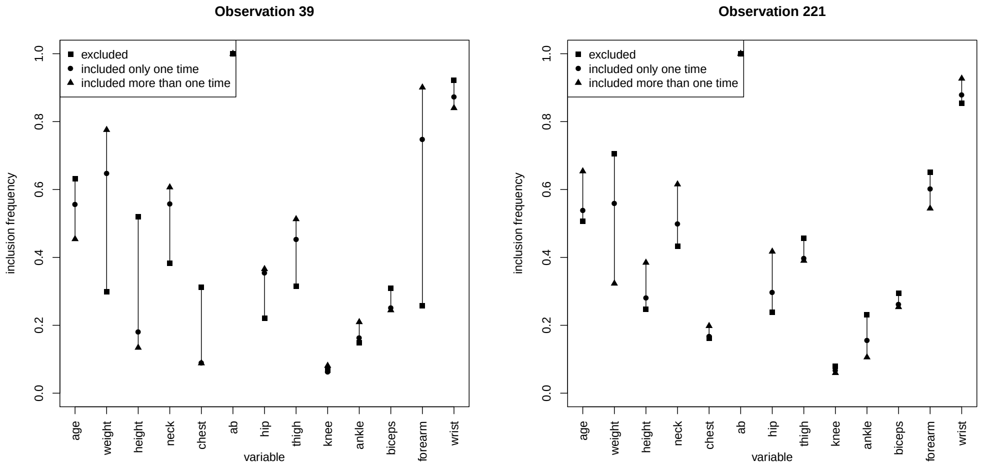


Figure 3: Body fat data: difference in the inclusion frequencies of the 13 variables when observation 39 (left graphic) or observation 221 (right graphic) are excluded (square), included one time (circle) or included more than once (triangle) in the pseudo-samples.

Figure 3 presents the aforementioned frequencies for the 2 possible influential points detected in the previous analyses, namely observations 39 and 221. We immediately notice the different strengths of their influence. Let us focus on the variables *weight* and *height*. The simple presence of observation 39 drastically changes the inclusion frequencies of the two variables, immediately flipping their ranks (when observation 39 is in the sample, *weight* has a higher inclusion frequency than *height*, when it is out of the sample, it is the other way around). A multiple presence of this observation in the sample does not really change the situation, and the differences between I-frequencies-1 and I-frequencies-M are minimal in comparison to the aforementioned differences between I-frequencies-1 and O-frequencies. On the contrary, observation 221 seems to have a much smaller effect. When it is included in the sample, the inclusion frequencies do not change so much (especially for *height*). To notice a certain effect, we should consider the I-frequencies-M. Only when observation 221 is included more than once in the bootstrap sample, indeed, the effect on the inclusion frequencies of *weight* and *height* is strong enough to flip their ranks (and, consequently, their relative chance to be included in the final model).

Note that this procedure allows us to turn into an advantage a possible pitfall of the bootstrap approach. The presence of duplicated observations, indeed, is usually considered a drawback, as pseudo-samples containing duplicated influential points may diverge from the original sample. Robust versions of the bootstrap procedure have been proposed to tackle this issue (see, e.g., Willems & Van Aelst, 2005). Here, instead, we take advantage of this characteristic of the bootstrap to highlight the effect of the possible influential points and to evaluate the strength of their effects. As one may argue that a fair comparison should involve only pseudo-samples without and with only one replication of a specific observation, in the following section we repeat the analyses using subsampling instead of bootrapping as a resampling technique. Pseudo-samples generated by subsampling, indeed, do not contain duplicated observations, as they are drawn without replacement.

4.1.4. Subsampling

As stated in Section 3.4, any resampling technique can be used to generate the pseudo-samples and consequently to compute the inclusion matrix. Figures A.3, A.4 and A.5 in the Web Appendix show the I-frequencies when using subsampling. As long as the size of the pseudo-samples (subsamples) is not extremely high, in this specific example we do not see any noticeable differences with the bootstrap approach (Figures A.3 and A.4). When the subsample size is too large, instead, the differences among samples got lost. Figure A.5, in particular, shows this situation in the case of subsamples of size $n - 2$. When choosing $n - 2$ as a subsample size, it may be more useful to use the O-frequencies instead of the I-frequencies in the analyses. When this is done, the effects of the single observations are taken to their extremes (see Figure A.6 in the Web Appendix). For example, in this dataset the O-frequencies for the variable *weight* are all equal or close to 1, but that related to observation 39, which is 0. For *height*, it is the other way around. On the one hand, having extreme differences among O-frequencies may help to graphically identify the possible influential point in an easier way; on the other hand, several points are equal to 0 or 1, and it is not possible to consider a Gaussian approximation for the distribution of the frequencies. As a consequence, the Grubbs' test cannot be applied. Note that the concept of O-frequencies computed in subsamples of size $n-2$ is very similar to the idea of delete-2 jackknife, that is sometimes used in the literature related for outlier detection (see, e.g., Martin et al., 2010).

4.1.5. Comparison with FSDA

Figure 4 reports two diagnostic plots obtained through the forward search approach to detect the effect of influential points on model selection by Atkinson & Riani (2002). Although this procedure mainly focuses on the effect of specific observations from a model point of view, rather than a variable point of view, it can be seen as an alternative to our approach, as suggested by a referee. The idea is rather simple: a model is first fitted on a carefully selected subsample of observations, and then stepwise re-fitted on the same subsamples enlarged, at each step, with one of the observation initially excluded (that closer to the fitted model). For details regarding the choice of the initial subsample and the ordering of the observations' inclusion we refer to the original paper. We just point out that the procedure is built so that the last observations are those with the highest probability to be influential points, being the farthest from the fitted model. Note that the method is robust against the masking effect (see Section 5 for more details), as weaker influential points are considered before the stronger ones.

The left plot of Figure 4, called "deletion statistics plot", shows the variables' importance depending on the subsamples. Each curve visualizes the evolution of the t-test for the nullity of a specific variable's regression coefficient when increasing the number of observations (the farthest observation from the fitted model is the last to be added). Concerning the strongest variables and the most influential observations the results are similar to those obtained with our approach and reported in Figure 1. The variable *ab* plays a predominant role, followed by *wrist*. Moreover, in the last step of the forward search, i.e. when the farthest observation, namely observation 39, enters in the sample, the curve related to *forearm* grows drastically, in accordance with the results reported in Figures 1 and 2 (and Table 1). Note that, in contrast to our approach, the deletion statistics plot is based on the full model,

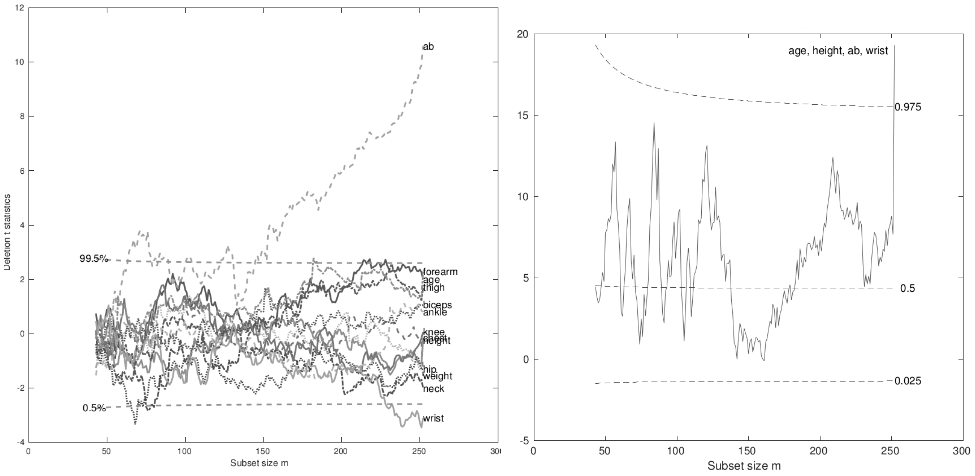


Figure 4: Body fat data: plots from function `FSRaddt` contained inside the MATLAB toolbox FSDA. Left: deletion statistics plot based on the full model; right: Mallow's C trajectory for the model including *age*, *height*, *ab* and *wrist*.

which is the starting point of our procedure but which is not selected in any replication.

The right plot of Figure 4, instead, focuses on the model which only includes *age*, *height*, *ab* and *wrist*, i.e. the model selected by backward elimination when observation 39 is excluded from the sample (see Table 1). In particular, the Mallow's C trajectory when increasing the sample size as described above is reported. As expected, this plot shows that the model describes very well the data until observation 39 is included. The Mallow's C trajectory, indeed, is within the 95% bands (dashed lines) until the very last step (i.e., inclusion of observation 39). This result is in line with ours, as we showed that, when observation 39 is in the sample, the selected model is not longer that including *age*, *height*, *ab* and *forearm* but that which includes *weight*, *ab*, *forearm* and *wrist* (see Table 1).

Summing up, in this example FSDA and our approach provide consistent results, as both identify observation 39 as a possible influential point. As mentioned above, however, the focus of the two approaches is relatively different: while the former considers the effect of a specific observation on the goodness of fit of the whole model, our approach focuses on the effect of possible influential points on the inclusion/exclusion of the single variables.

4.2. Myeloma data

In this section we investigate the second dataset. We report only the results obtained when using bootstrap as a resampling technique. As we can see in Figure 5, it seems that there are several points separated from the others. This is a consequence of the small sample size (there are only 48 events). In this situation, especially in a survival context, each observation noticeably influences the model building procedure. To check if these points are influential points, it may be better to rely on the standardized I-frequencies and to compare them with the rejection region of the Grubbs' test.

We report the results in Figure 6. When we consider the rejection region of the Grubbs' test, it is clear that there are no strong influential points in this dataset. The only point that is inside the rejection region and relatively far from the boundary is the largest I-frequency for variable *protein*. This I-frequency is that computed on only pseudo-samples including observation 44. Please note that another point related to this observation is inside the rejection region, namely, the smallest I-frequency for the variable *hemoglob*.

As for the other dataset, we can deepen the analysis on the effect of this observation by comparing the inclusion frequencies computed in pseudo-samples without it (O-frequencies), in pseudo-samples in which it appears only once (I-frequencies-1) and in pseudo-samples in which it appears more than one time (I-frequencies-M). The results are shown in Figure 7. As expected, the most relevant effect is related to variable *protein*. The inclusion of observation 44 in the samples increases the times in which this variable is included in the model. This effect is stronger when the observation is included more than once. In contrast, the presence of observation 44 decreases the inclusion frequency of the variable *hemoglob*. In this case, there is no strong difference whether the observation is included once or more than one time.

As we noted above, in this dataset the effective sample size is quite small and every non-censored observation may have a relatively strong influence on the model-building process. Observation 44 stands out as the most influential one, but there is no strong evidence (e.g., large distance from the 99% rejection region boundaries of the Grubbs' test) that suggests us that it is an outlier. Its presence in the rejection region for variables *hemoglob* and *protein* may simply be related to the type I error of the Grubbs' test.

The small sample size also influenced our choice of the bootstrap technique. In order to include an acceptable number of events (here, in particular, the same amount of the original sample) in all pseudo-samples, we

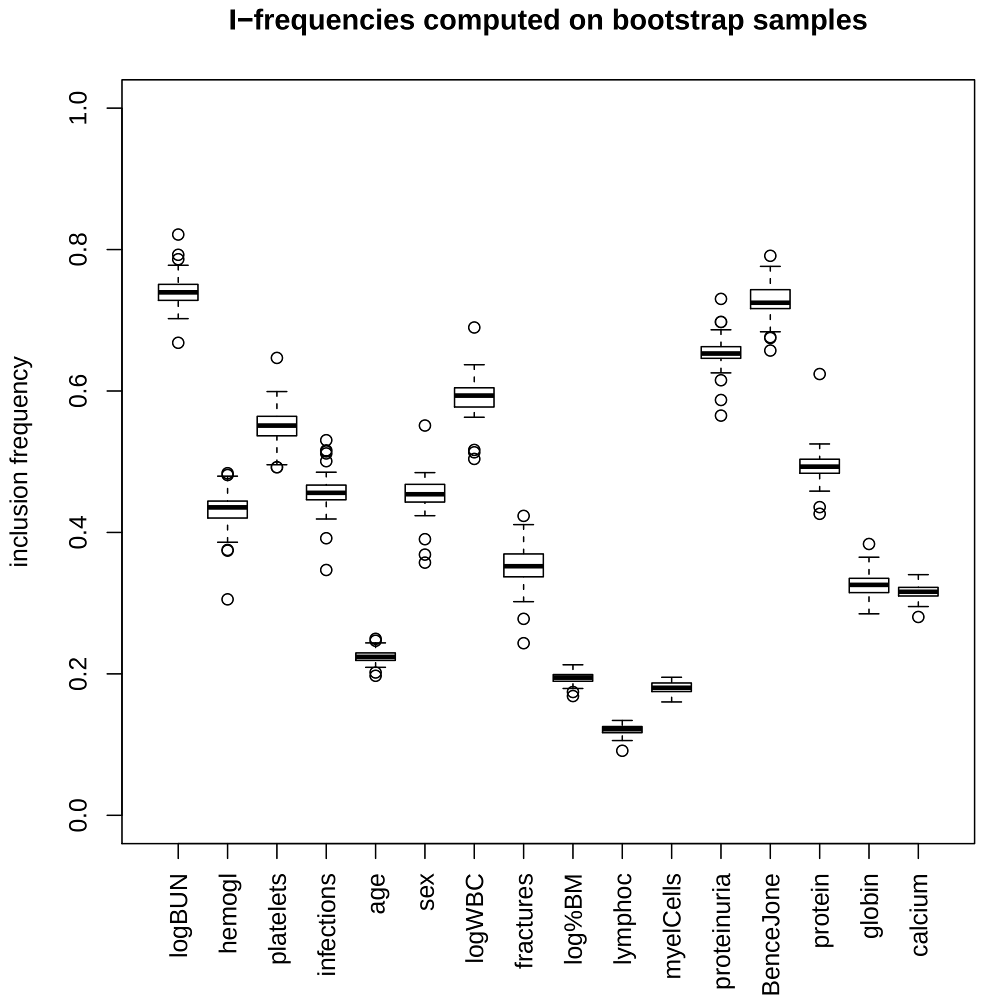


Figure 5: Myeloma data: boxplots that summarize the I-frequencies obtained for the 16 variables, based on boostrap pseudo-samples.

implemented a stratified bootstrap, i.e., we resampled separately censored and non-censored observations.

5. Discussion

In this paper we showed how the information present in the inclusion matrix can be used to identify possible influential points. Provided that a

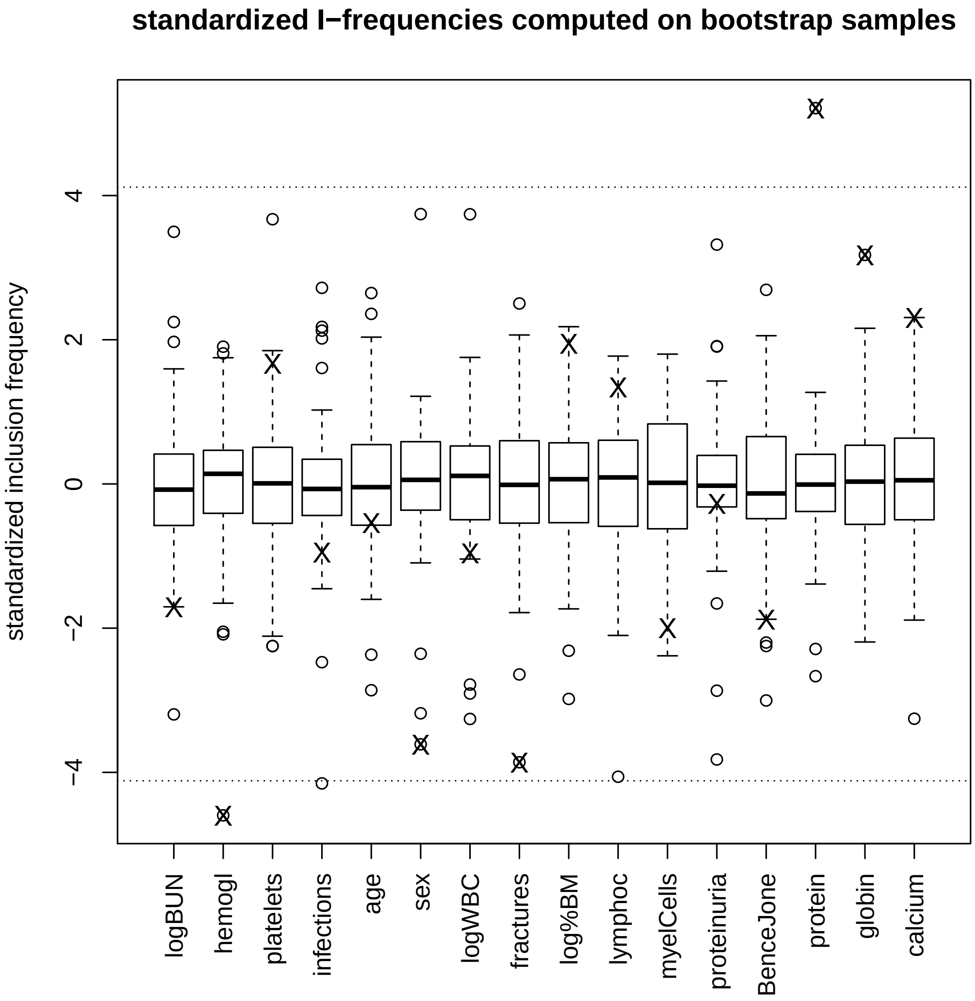


Figure 6: Myeloma data: boxplots that summarize the standardized I-frequencies (computed on bootstrap pseudo-samples) obtained for the 16 variables. The symbol "X" denotes the observed values of the test statistics for observation 44. The dashed lines delimit the 99% rejection region of a Grubbs' test (including a correction for the multiplicity of the tests).

resampling based approach is used to select a model, to assess model stability or to use model averaging for the derivation of a predictor, the generation of the inclusion matrix is a step not requiring any further computation. It uses available information.

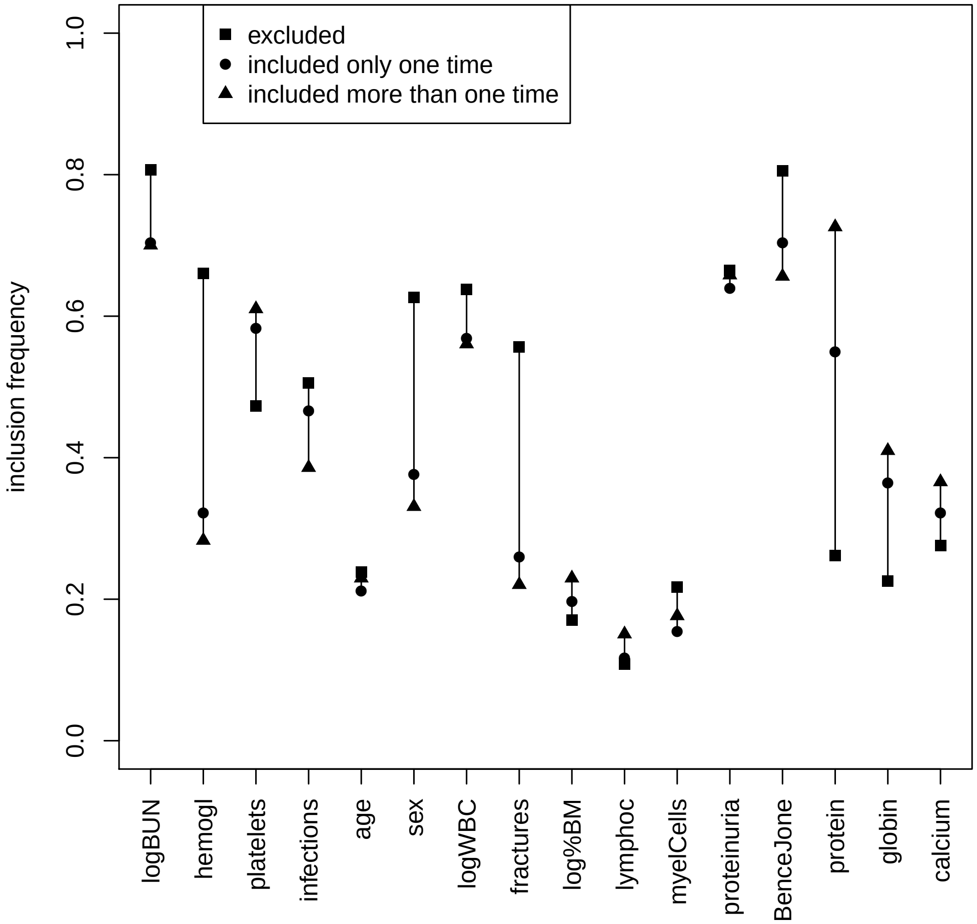


Figure 7: Myeloma data: difference in the inclusion frequencies of the 16 variables when observation 44 is excluded (square), included one time (circle) or included more than once (triangle) in the pseudo-samples.

Another advantage of our approach is the possibility of having a clear graphical visualization of the results, which allows the user to easily spot possible influential points. As stated in the literature, the identification of an influential point should be done by the user, and not fully delegated to an automatic procedure (Billor et al., 2000).

We considered graphical inspections based both on the simple I-frequencies and on their standardized version. As mentioned before, both approaches have advantages and disadvantages. In particular, when using the simple I-frequencies, we can also have an impression, in the same graphics, of the importance of the variables whose inclusion frequencies may be influenced by a specific observation. As we saw in the first example, this allows us to focus on points of interest, namely influential points which change the selection of the final model, and avoid the investigation of observations that influence the inclusion frequencies of variables which will not be included anyhow. Moreover, the plot of the simple I-frequencies also gives an idea on the variance of their inclusion frequencies, which may be useful in the model building process.

On the other hand, the use of standardized I-frequencies allows us to have a better insight into the influence of the single points. The values of the standardized I-frequencies, indeed, can be contrasted to the rejection region of an univariate test for outliers, in our paper we use the Grubbs' test, to have a more objective estimates of their influence. In studies with very small sample sizes (see the myeloma data) we may identify several observations which seem to be critical. A final assessment needs to consider further criteria from the study.

In this paper we did not consider in detail the problem of the masking effect (Bendre & Kale, 1985). The presence of a strong influential point, indeed, may hide the effect of an observation that has a smaller but still significant influence in the opposite direction. While there are methods, as that by Atkinson & Riani (2002) shown in Section 4.1.5, constructed with this in mind, ours is not specifically designed to tackle the masking effect issue. Nevertheless, we note that in the body fat data example our method was able to detect observation 221 as a possible influential point despite the fact that its effect is the opposite to that of observation 39. Observation 221 has never been identified as an influential point, probably because, in the previous studies, methods that do not take into account the masking effect have been implemented. Our methods, instead, seems to be less prone to suffer from this specific issue.

Acknowledgments

We thank Sarah Tegenfeldt for language improvements and the two anonymous referees for their comments, which led to an improved version of the paper. DBR was partially supported by grant BO3139/4-2 to ALB and WS was supported by grant SA580/8-2. Both grants are from the German Research Foundation (DFG).

Bibliography

Atkinson, A. C., & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer, New York.

Atkinson, A. C., & Riani, M. (2002). Forward search added-variable t-tests and the effect of masked outliers on model selection. *Biometrika*, (pp. 939–946).

Augustin, N., Sauerbrei, W., & Schumacher, M. (2005). The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling*, *5*, 95–118.

Bendre, S., & Kale, B. (1985). Masking effect on tests for outliers in exponential models. *Journal of the American Statistical Association*, *80*, 1020–1025.

Billor, N., Hadi, A. S., & Velleman, P. F. (2000). Bacon: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, *34*, 279–298.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, *53*, 603–618.

Chen, C.-H., & George, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine*, *4*, 39–46.

Chen, C.-H., & Wang, P. C. (1991). Diagnostic plots in Cox's regression model. *Biometrics*, *47*, 841–850.

De Bin, R., Janitza, S., Sauerbrei, W., & Boulesteix, A.-L. (2016). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometics*, *72*, 272–280.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, *21*, 488–506.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, *23*, 881–890.

Gong, G. (1982). Some ideas on using the bootstrap in assessing model variability. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface* (pp. 169–173). Springer New York.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, *21*, 27–58.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*, 1175–1189.

Hansen, B. E., & Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, *167*, 38–46.

Hartigan, J. A. (1969). Using subsample values as typical values. *Journal of the American Statistical Association*, *64*, 1303–1317.

Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*, 879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, *14*, 382–401.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, *4*, 265–266.

Krall, J. M., Uthoff, V. A., & Harley, J. B. (1975). A step-up procedure for selecting variables associated with survival. *Biometrics*, *31*, 49–57.

Kuk, A. Y. (1984). All subsets regression in a proportional hazards model. *Biometrika*, *71*, 587–592.

Martin, M. A., Roberts, S., & Zheng, L. (2010). Delete-2 and delete-3 jackknife procedures for unmasking in regression. *Australian & New Zealand Journal of Statistics*, *52*, 45–60.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*, 417–473.

Myint, P. K., Kwok, C. S., Luben, R. N., Wareham, N. J., & Khaw, K.-T. (2014). Body fat percentage, body mass index and waist-to-hip ratio as predictors of mortality and cardiovascular disease. *Heart*, *100*, 1613–1619.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `https://www.R-project.org/`.

Riani, M., Perrotta, D., & Torti, F. (2012). FSDA: a MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, *116*, 17–32.

Rospleszcz, S., Janitza, S., & Boulesteix, A.-L. (2016). Categorical variables with many categories are preferentially selected in bootstrap-based model selection procedures for multivariable regression models. *Biometrical Journal*, *58*, 652–673.

Royston, P., & Sauerbrei, W. (2007). Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. *Computational Statistics & Data Analysis*, *51*, 4240–4253.

Royston, P., & Sauerbrei, W. (2008). *Multivariable Model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*. Wiley, Chichester.

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L., & Binder, H. (2015). On stability issues in deriving multivariable regression models. *Biometrical Journal*, *57*, 531–555.

Sauerbrei, W., & Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine*, *11*, 2093–2109.

Srivastava, M. (1980). Effect of equicorrelation in detecting a spurious observation. *Canadian Journal of Statistics*, *8*, 249–251.

Su, X., & Tsai, C.-L. (2011). Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*, 261–268.

Tsao, M., & Ling, X. (2012). Subsampling method for robust estimation of regression models. *Open Journal of Statistics*, *2*, 281.

Wan, A. T., Zhang, X., & Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, *156*, 277–283.

Wang, H., Zhang, X., & Zou, G. (2009). Frequentist model averaging estimation: a review. *Journal of Systems Science and Complexity*, *22*, 732–748.

Willems, G., & Van Aelst, S. (2005). Fast and robust bootstrap for lts. *Computational Statistics & Data Analysis*, *48*, 703–715.

Detection of influential points as a byproduct of resampling-based variable selection procedures

Riccardo De Bin

Department of Mathematics, University of Oslo
Department of Medical Informatics, Biometry and Epidemiology, University of Munich

Anne-Laure Boulesteix

Department of Medical Informatics, Biometry and Epidemiology, University of Munich

Willi Sauerbrei

Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg

Email addresses: debin@math.uio.no, postal address: Postboks 1053 Blindern 0316 Oslo (Norway), telephone:(+47)22855859 (Riccardo De Bin), boulesteix@ibe.med.uni-muenchen.de (Anne-Laure Boulesteix), wfs@imbi.uni-freiburg.de (Willi Sauerbrei)

Detection of influential points as a byproduct of resampling-based variable selection procedures

Riccardo De Bin

Department of Mathematics, University of Oslo
Department of Medical Informatics, Biometry and Epidemiology, University of Munich

Anne-Laure Boulesteix

Department of Medical Informatics, Biometry and Epidemiology, University of Munich

Willi Sauerbrei

Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg

Abstract

Influential points can cause severe problems when deriving a multivariable regression model. A novel approach to check for such points is proposed, based on the variable inclusion matrix, a simple way to summarize results from resampling-based variable selection procedures. These procedures rely on the variable inclusion matrix, which reports whether a variable (column) is included in a regression model fitted on a pseudo-sample (row) generated from the original data (e.g., bootstrap sample or subsample). The variable inclusion matrix is used to study the variable selection stability, to derive weights for model averaged predictors and in others investigations. Concentrating on variable selection, it also allows understanding whether the presence of a specific observation has an influence on the selection of a variable. From the variable inclusion matrix, indeed, the inclusion frequency (I-frequency) of each variable can be computed only in the pseudo-samples (i.e., rows) which contain the specific observation. When the procedure is repeated for each

Email addresses: `debin@math.uio.no, postal address: Postboks 1053 Blindern 0316 Oslo (Norway), telephone:(+47)22855859` (Riccardo De Bin), `boulesteix@ibe.med.uni-muenchen.de` (Anne-Laure Boulesteix), `wfs@imbi.uni-freiburg.de` (Willi Sauerbrei)

observation, it is possible to check for influential points through the distribution of the I-frequencies, visualized in a boxplot, or through a Grubbs' test. Outlying values in the former case and significant results in the latter point to observations having an influence on the selection of a specific variable and therefore on the finally selected model. This novel approach is illustrated in two real data examples.

Keywords: bootstrap; Grubbs' test; inclusion frequency; model averaging; outliers; subsampling.

1. Introduction

In the construction of a statistical model, an important aspect to take into consideration is its stability. It is well known, indeed, that small perturbations in the data may lead to the selection of different models. For example, several papers show that variable selection procedures, such as backward elimination or forward selection, may provide very different sets of relevant variables, and consequently very different models, when applied to different bootstrap samples generated from the same dataset (Sauerbrei et al., 2015).

In the literature, different approaches have been proposed to handle this issue. From a variable point of view, resampling-based variable selection techniques can handle the instability issue by investigating the inclusion frequencies of the single variables (Gong, 1982; Chen & George, 1985). The idea is rather simple. Several pseudo-samples are generated via a resampling technique and a variable selection procedure is applied to select the best model in each of them. The proportion of models which contain the specific variable (inclusion frequency) is used as an indicator of the importance of the variable itself, and those variables with higher inclusion frequencies are used in the final model.

From a model point of view, model averaging is a technique which aims to deal with model uncertainty by fitting different models on the data and then summarizing their results. For example, in linear regression, a regression coefficient is estimated as a weighted mean of the corresponding estimates computed in each model. In particular, in the resampling-based approaches

Supplementary Material and the R-code to reproduce the results are available in a Web Appendix.

the weights are obtained by generating several pseudo-samples via a resampling technique and evaluating for how many of these pseudo-samples the different models are selected by a variable selection procedure. Other kinds of weights are based on information criteria, Mallows' criterion, etc. For a review on model averaging and on the different alternatives for the computation of the weights, we refer the reader to Wang et al. (2009). That paper, in particular, considers the frequentist approach. For a review about Bayesian model averaging, a classical reference is Hoeting et al. (1999).

Both resampling-based variable selection and resampling-based weights for model averaging require the application of a variable selection technique to several pseudo-samples. The goal of this paper is to show that the information collected in this part of the analysis can be used to check for influential points, such as outliers or single observations that have a high impact on the results. It is well known that influential points can cause problems when selecting a statistical model. For example, the inclusion or exclusion of a single or a few observations can have a dramatic effect on variables selected and on the issue of selecting linear or nonlinear function for a continuous variable (Royston & Sauerbrei, 2007). The literature on influential point detection is vast, and countless approaches have been proposed. For a simple and concise overview we refer the reader to Su & Tsai (2011) and references therein.

The detection of influential points as a byproduct of model-building procedures is not new. Tsao & Ling (2012), for example, exclude from the final model fitting procedure those observations that are not included in any of the pseudo-samples that lead to good models in terms of goodness-of-fit. A similar approach is used by Sauerbrei et al. (2015), who consider the selection probabilities of some "best models" and identify as influential points those observations which are able to modify these selection probabilities. Both approaches handle the influential point detection issue from a model point of view, ignoring the effect of these observations on the single variables. In this paper we consider the problem from a variable point of view, though maintaining a multivariable approach.

Finally, we mention Atkinson & Riani (2002), who also studied the effect of influential points from a model building point of view, using a forward search procedure (Atkinson & Riani, 2000, Ch. 2). We contrast our and their approaches in Section 4.1.5.

The paper is structured as follows. Section 2 presents two datasets later used as real examples. A brief introduction to model averaging and resampling-based variable selection is presented in Section 3, together with the description of our approach. The application of the method to the data is reported in Section 4. Finally, Section 5 contains a short discussion.

2. Data

2.1. Body fat data

The estimate of the percentage of body fat is considered a good indicator to assess the health of patients (see, e.g., Myint et al., 2014). Johnson (1996) presents a dataset in which the percentage of body fat (PBF) is collected from 252 men, together with the information about 13 further quantities, namely *age*, *weight*, *height* and 10 continuous body circumference measurements that are considered variables with potential influence on PBF. The data are publicly available at `http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book/Multivariable_Model-building/downloads/datasets/edu_bodyfat_both.zip`.

	BIC		$\alpha = 0.05$		AIC	
variable	in	out	in	out	in	out
age		✓		✓	✓	✓
weight	✓		✓		✓	
height		✓		✓		✓
neck					✓	✓
chest						✓
ab	✓	✓	✓	✓	✓	✓
hip					✓	
thigh					✓	
knee						
ankle						
biceps						
forearm	✓		✓		✓	✓
wrist	✓	✓	✓	✓	✓	✓

Table 1: Body fat data: result of a backward elimination procedure using three different selection criteria (BIC, significance level 0.05, AIC), with (in) and without (out) observation 39.

It is important to note that this dataset contains at least one influential point. Royston & Sauerbrei (2007), in particular, show that observation 39 highly influences the choice of the fractional polynomial function used to model the relationship between outcome and variables. Although some variables seem to have a non-linear effects on the outcome, we re-analyse this dataset under the assumption of linear effects. Non-linear effects are not that strong and this simplifying assumption seems acceptable for the main purpose of this paper.

To show the effect of observation 39 in a classical model-building procedure, we report in Table 1 the models obtained with backward elimination when this observation is included/excluded from the sample. Three common inclusion criteria are used. In this example, results are identical for BIC and $\alpha = 0.05$. As commonly seen in the literature (see, e.g., Sauerbrei et al., 2015), more variables are selected with AIC. We note that the presence/absence of observation 39 in the sample leads to substantially different models. The selections of variables *age*, *weight*, *height* and *forearm* are clearly affected.

2.2. Myeloma data

As an application of our method to a different kind of outcome, we also use a dataset with a time-to-event outcome. In particular, we consider a study on patients with multiple myeloma presented by Krall et al. (1975), in which the outcome is the survival time of the patients. The 16 variables are either binary or continuous. We consider the proportional hazard assumption acceptable, being this dataset analyzed several times in the literature by using the Cox model (see, e.g., Kuk, 1984; Chen & Wang, 1991). The sample size is small, consisting of 65 patients with 48 events. As for the body fat data, we use the simplifying assumption that the effect of continuous variables is linear. This dataset is also publicly available on the same website (`http://.../myeloma.zip`).

3. Methods

3.1. Resampling-based variable selection

One aim of a resampling-based variable selection is to select the relevant variables to include into a statistical model in a robust way, with the idea that the same model should be identified despite small perturbations in the data. In practice, a resampling technique, such as bootstrap or subsampling, is applied to the original dataset to generate several pseudo-samples, in order to mimic small perturbations in the data. As a sample with (bootstrap) or without (resampling) replacement from the original dataset, indeed, these pseudo-samples can be considered new instances of the data-generating mechanism, similar but not identical to the observed one. A variable selection technique, for example backward elimination, is then applied to each pseudo-sample. The proportion of pseudo-samples in which each variable is selected is called "inclusion frequency" and it is used to discriminate between relevant and irrelevant variables. The variables with higher inclusion frequencies are included in the final model, while the others are discarded. Table 2 reports an example of the computation of the inclusion frequencies. For further details and approaches to handle issues related to the dependence of inclusion frequencies among pairs of variables, see Sauerbrei & Schumacher (1992).

	variable							
pseudo-sample	V_1	V_2	V_3	$\dots$	V_{q-1}	V_q		model
1	1	0	1	$\dots$	0	1	→	$\mathcal{M}_1$
2	0	1	1	$\dots$	0	0	→	$\mathcal{M}_2$
3	1	0	1	$\dots$	0	1	→	$\mathcal{M}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	→	$\vdots$
B	1	0	1	$\dots$	0	0	→	$\mathcal{M}_k$
inclusion frequency	0.961	0.243	1.000	$\dots$	0.000	0.693		

Table 2: Illustration of a variable inclusion matrix. It can be used to compute the resampling-based weights in a model averaging procedure (last column) or to compute the variable inclusion frequencies in a resampling-based variable selection procedure (last row).

3.2. Model averaging with resampling-based weights

The idea of model averaging consists in making inference on a parameter of interest by using several models instead of a single one. Consider K models $\mathcal{M}_1, \dots, \mathcal{M}_K$. The parameter estimate $\hat{\theta}$ is defined as the weighted average of the estimates computed across the K model ($\hat{\theta}_{\mathcal{M}_k}$), in formula

$$\hat{\theta} = \sum_{k=1}^{K} w_k \hat{\theta}_{\mathcal{M}_k}. \quad (1)$$

A highly relevant point is the choice of the weights w_k. In the literature several procedures have been proposed, for example based on information criteria (e.g. Buckland et al., 1997; Hjort & Claeskens, 2003) or Mallows' criterion (e.g. Hansen, 2007; Wan et al., 2010). Here we focus on weights based on a resampling approach, such as in, among others, Buckland et al. (1997); Augustin et al. (2005). As for resampling-based variable selection, a large number B of pseudo-samples are generated through a resampling technique and, to each pseudo-sample, a variable selection procedure is applied. In contrast to the previous approach, here the focus is not on the variables but on the resulting models. The proportion of time in which the model $\mathcal{M}_k$ is selected gives, for $k = 1, \ldots, K$, the weight w_k,

$$w_k = \frac{\#\mathcal{M}_k}{B}.$$

These weights are then used in formula (1). Although the inclusion matrix is the same as before (see Table 2), now the information is extracted on the direction of the rows (models).

Note that Hansen & Racine (2012) also used a resampling technique (in their case, jackknife) to derive the weights. Nevertheless, their approach relies on the estimate of the mean square error and therefore is theoretically different from the procedure described above.

3.3. Detection of possible influential points

3.3.1. From the inclusion matrix to the frequency matrix

We saw that both resampling-based variable selection and model averaging with resampling-based weights rely on an inclusion matrix. In each row, this matrix provides the information about which variables are included in the best model fitted on that particular pseudo-sample. For example, in a study with q variables, each row of the inclusion matrix is a q-dimensional vector containing 0 (variable not included) and 1 (variable included). The number of rows is arbitrary, and corresponds to the number of iterations performed. Table 2 reports an illustration of an inclusion matrix. As we saw above, in a resampling-based variable selection procedure this matrix is used to compute the inclusion frequencies for the variables (column averages), in a model averaging procedure to compute the weights (each row corresponds to a model).

Since each row corresponds to a pseudo-sample, the inclusion matrix also provides us with important information about the relationship between variables and observations. In addition to the inclusion/exclusion of the variables in the selected model, indeed, for each row we know which observations belong to the particular pseudo-sample and which do not. Combining these two aspects, we can evaluate the effect of a specific observation on the inclusion frequencies of the variables. For each observation i, we can estimate inclusion frequencies of all variables separately for samples including or excluding i. For a variable V_j, the two frequencies should be similar if the observation i has no effect on its inclusion and different if i has an influence on the inclusion of V_j.

Let us focus on the inclusion frequencies obtained by considering only the pseudo-samples in which a specific observation is included. For each observation $i = 1, \ldots, n$, we compute these inclusion frequencies (hereafter, "I-frequencies", where "I" stands for "in") for all variables, obtaining a q-dimensional vector in which each entry corresponds to one variable (q is the number of variables). By merging these vectors, we obtain a $n \times q$ matrix of I-frequencies (hereafter, "I-frequency matrix"), as that reported in Table 3. In this example, in the pseudo-samples in which observation x_1 is included (first row), the variable V_1 is selected 0.969 of the times, V_2 0.015, and so on.

observation	variable					
included	V_1	V_2	V_3	$\ldots$	V_{q-1}	V_q
1	0.969	0.015	0.553	$\ldots$	0.000	0.292
2	1.000	0.030	0.492	$\ldots$	0.000	0.376
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$n-1$	1.000	0.015	0.603	$\ldots$	0.000	0.361
n	0.984	0.092	0.569	$\ldots$	0.000	0.276

Table 3: Illustration of a I-frequency matrix. For each variable (column), it reports its I-frequencies, i.e. the inclusion frequency computed only on pseudo-samples in which a specific observation (row) is included.

3.3.2. I-frequency matrix and detection of influential points

If there is no influential point in the sample, we expect the values in the column of the I-frequency matrix to be very similar to each other. Conversely, the effect of an influential point would be visible in values that are strongly separated from the rest. Let us consider, as an example, an influential point, let say x_i, which strongly influences the significance of a variable V_j, in the sense that it forces V_j to enter into the model. Focusing on the column related to V_j, we would expect in the i-th row of the I-frequency-matrix a value much larger than all other values present in the same column.

Visualization. The easiest way to identify possible influential points is to plot the column values of the I-frequency matrix in boxplots, and take advantage of what Friedman and Tukey call "the human gift for pattern recognition" (Friedman & Tukey, 1974). The boxplot is a simple and effective tool to display the I-frequencies of a variable and to identify those that are far from the median value. In particular, in the standard way of drawing a boxplot, the extreme observations are not included in the whiskers and are plotted as separated points. ~~For example, in R (R Core Team, 2016), this happens~~ Usually, this is done for points farther than 1.5 times the interquartile range from the first/third quartile. The farthest points are the values we are interested in, because they represent the most anomalous inclusion frequencies. One can then easily go back to the frequency matrix and identify the rows which correspond to these values, and, consequently, which are the possible influential points. In the case of no influential points, instead, we would expect no strongly separated points, i.e. a plot in which all values would be included or would be close to the boxplot's whiskers. Note, however, that identifying possible outliers among the points outside the whiskers is a delicate task, and more objective criteria may be necessary (see also Section 3.3.3).

Remark. The column variance of the I-frequency matrix can also be seen as an indicator of the "trustworthiness" of the variable inclusion frequency. Smaller variance, indeed, means an inclusion frequency that does not change too much in the case of small perturbations in the data. If for any reason we are in doubt whether a variable should or should not be included in the model, the variance may be a further argument to support our choice. For example, in the case of two correlated variables with similar inclusion frequencies, we may prefer to select that for which we obtain a smaller variance, because less influenced by small perturbations in the data.

3.3.3. Grubbs' tests

Although several researchers advocate graphical investigations to detect influential points, in some extends it may be advantageous to rely on a statistical test. From our point of view, we need to test whether the most extreme (i.e., farthest from the median frequency) I-frequency is an outlier for each variable. In the case of a positive answer, it would mean that one single observation, let us say $x_{(n)}$, is able to change the inclusion or exclusion of a variable in the model in a significant way. In other words, that $x_{(n)}$ may be an influential point. In order to evaluate the influence of each observation on each variable, we analyze the I-frequency matrix column by column. In this way, we can simply apply to each column a simple univariate test, such as the Dixon's Q (Dixon, 1950) and the Grubbs' G (Grubbs, 1950). Due to the dependence of the former to the sample size, here we use the latter. It is worth stressing, in any case, that our analysis is meant as explorative. Once the aforementioned $x_{(n)}$ has been selected by our procedure, it is the responsibility of the practitioner to evaluate the exact nature of the observation (i.e., whether it is actually an influential point).

Given a sample $x_1, \dots, x_n$ from a Gaussian distribution, the Grubbs' test rejects the null hypothesis, defined as the absence of outliers, if

$$\max_{i=1,\dots,n} \frac{|x_i - \bar{x}|}{s} > C(\alpha, n) = (n-1)\sqrt{\frac{t^2_{1-\alpha/(2n),n-2}}{(n-2+t^2_{\alpha/(2n),n-2})}},$$

where $\bar{x}$ denotes the sample mean, s the estimated standard deviation and $t_{1-\alpha/(2n),n-2}$ the quantile $1-\alpha/(2n)$ of a t distribution with $n-2$ degrees of freedom. Here α is the significance level on which the test is conducted; since we repeat the test for each variable, it may be necessary to implement a correction for the multiplicity of the tests.

Visualization. For an easy identification of the influential points, it may be convenient to visualize the results in a graphic. Our suggestion is to plot, for each variable (i.e., for each column of the I-frequency matrix), the standardized I-frequency. This value is strictly related to the test statistic of the Grubbs' test, with the difference that we do not consider the absolute value but simply the difference between the value and the mean. If one value is outside the bands $\pm C(\alpha, n)$ it means that the I-frequency is an outlier and the corresponding observation may be an influential point. Please note that the Grubbs' test is constructed to identify the presence of one outlier. In general, a new critical value $C(\alpha, n)$ should be considered in the case of multiple outliers, namely

$$C(\alpha, n, k) = (n-k)\sqrt{\frac{t^2_{1-\alpha/(2(n-k+1)),n-k-1}}{(n-k-1+t^2_{\alpha/(2(n-k+1)),n-k-1})}},$$

where k indicates the number of outliers whose presence in the sample one wants to test. Nevertheless, for reasonably large sample size ($n > 50$), the critical value does not change much with k and the original $C(\alpha, n)$ can be used.

Remark. Note that the I-frequencies do not follow a Gaussian distribution, which is an assumption of the Grubbs' test. Their distribution may be better described by a beta distribution with accumulation points on the boundaries (0 and 1). Nevertheless, the beta distribution can be approximated by a Gaussian distribution when its coefficients are sufficiently large, i.e., when the data points are far from the boundaries. In fact, we are only interested in these cases. I-frequencies close to 0, indeed, are related to irrelevant variables, which should not be included into the final model. On the other extreme, I-frequencies close to 1 are typical of strong variables, which are almost always included in the model. In these two cases, the possible presence of an influential point would not change our decision to include or exclude the variable from the final model. In contrast, the dependence among the I-frequencies, which are computed on the same pseudo-samples, is not a problem. It has been shown that Grubbs' test is robust against deviation from independence (Srivastava, 1980).

3.4. Effect of the choice of the resampling technique

The construction of the inclusion matrix needs the implementation of a resampling technique to generate the pseudo-samples. Historically, bootstrap (Efron, 1979) has been the most used approach. It generates the pseudo-samples by sampling with replacement n observations (i.e. the original sample size) from the original sample. Alternatives such as subsampling (Hartigan, 1969) have been also considered (see, e.g. Meinshausen & Bühlmann, 2010). Subsampling consists of sampling without replacement $m < n$ observations from the original sample.

The choice of the resampling technique should be driven by considerations on the model-building procedure, both for model averaging and for resampling-based variable selection (for a recent study in the latter case, see De Bin et al., 2016). For example, when there are variables with different numbers of categories, the use of the bootstrap may cause misleading results (Rospleszcz et al., 2016). However, from an influential point detection point-of-view, the bootstrap gives the possibility to separately consider pseudo-samples by the number of times (e.g.. 0, 1, more than 1) they include an observation i (Royston & Sauerbrei, 2008, Section 8.5.1). For this reason, we consider the bootstrap in our study.

3.5. Software

The following analysis have been computed using R (R Core Team, 2016) and ~~Matlab~~the FSDA MATLAB toolbox (Riani et al., 2012). The R-code implementing the analyses is reported in the Web Appendix and the data are publicly available. That allows reproducibility of our study. Moreover, we plan to upload soon a specific R-package to the CRAN.

4. Results

4.1. Body fat data

From the original body fat sample, we generate 2000 bootstrap samples. To these pseudo-samples we applied a backward elimination procedure with significance level $\alpha = 0.05$. As a result, we obtain a 2000×13 inclusion matrix. We explained that this matrix can be used to perform variable selection or to compute the weights for a model averaging procedure. Here, instead, we use it to generate the I-frequency matrix (as Table 3) and to check the possible presence of influential points.

4.1.1. I-frequency matrix

The I-frequency matrix is a 252×13 matrix whose columns report the I-frequencies for the 13 variables and whose rows correspond to the observations as explained above.

Figure 1 shows boxplots of the inclusion frequencies (see Section 3.3.1). We note some points that are far from the respective median frequencies and, in general, from all other I-frequencies. This fact is a sign of the possible presence of influential points in the data. In particular, for the variables *weight*, *height*, *chest* and *forearm* we note four points (one for each variable) with this characteristic. All four points correspond to the inclusion frequencies computed on pseudo-samples which include observation 39. As stated in Section 2.1, observation 39 is a known outlier, and our method is able to correctly identify it. Interestingly, there seems to be an observation (observation 221) which also has an effect on the inclusion of *weight*, in an opposite direction than observation 39.

Figure 1 contains further information. First of all, we can see the effect of the influential point in the model building process. It is clear that observation

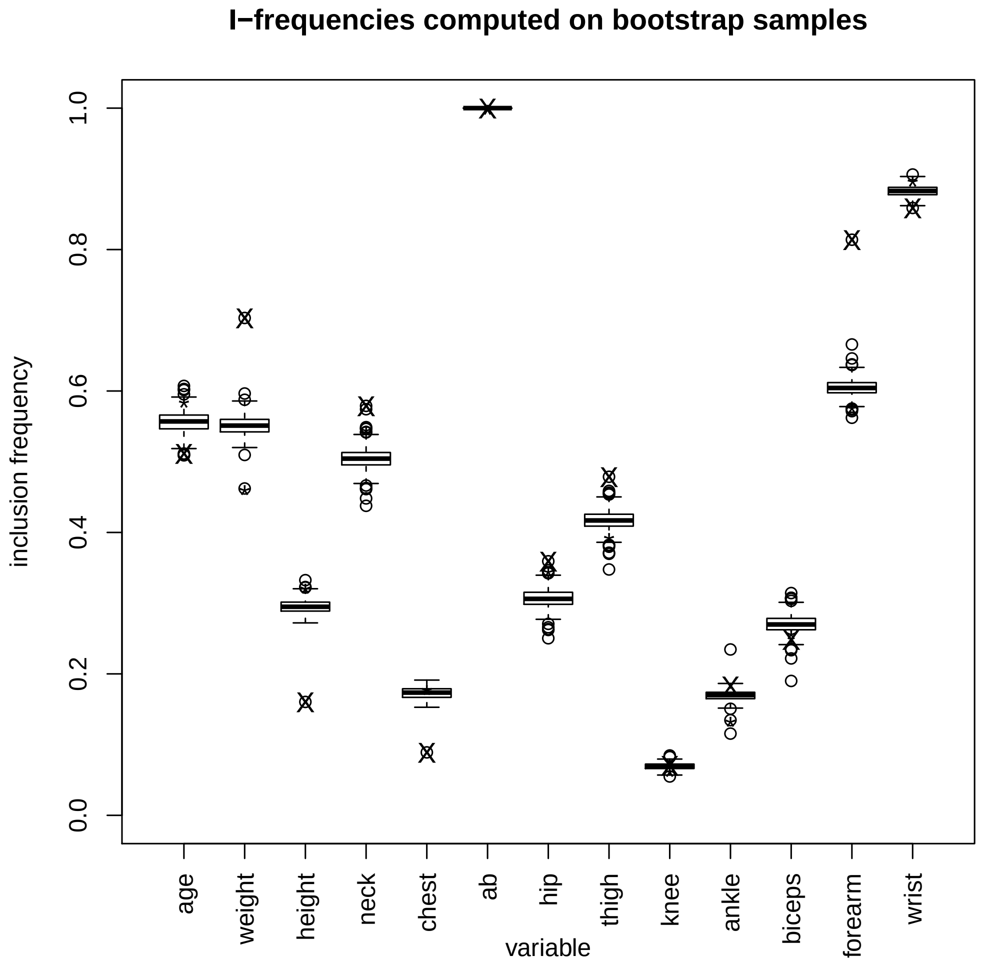


Figure 1: Body fat data: boxplots that summarize the I-frequencies obtained for the 13 variables. The symbols "X" and "*" denote the I-frequencies obtained in pseudo-samples (here generated by using bootstrap) which include observation 39 and 221, respectively.

39 leads to an overestimation of the importance of *weight* and *forearm*, and an underestimation of *height* and *chest*. Secondly, we can visualize the strengths of the variables. We note, for example, that *ab* (abdominal circumference) is a very strong variable, which is always included in the model, while *knee*, which is included in only a few models (low inclusion frequency), seems irrelevant. Moreover, this figure allows us to compare the variances of the inclusion

frequencies. As mentioned before (Section 3.3.2), this information may be useful in the model building procedure.

4.1.2. Grubbs' tests

As described in Section 3.3.3, we can perform univariate tests on the I-frequencies for each variable. To visualize the results of the tests, we show boxplots of the standardized I-frequencies (observed test statistics) and draw the rejection region for the Grubbs' test with a significance level of 99% (Figure 2).

It is fairly simple to detect possible outliers. Note that the observed values for the test statistics for observation 39 are very improbable under the null hypothesis (no outlier), i.e., the points are deeply inside the rejection region and very far from its boundaries. This is true in particular when considering the variables *weight*, *height* and *forearm*. Moreover, observation 221 shows potential to be an influential point, with effect on the selection of *weight*. Although the corresponding point is not so distant from the rejection region boundary, we should take into consideration that its position is influenced by that of observation 39, which has opposite effect. Some pseudo-samples which include observation 221 also contain observation 39, and tend to increase the value of this point. Despite the masking effect of observation 39, observation 221 is inside the rejection region.

Further points are inside the rejection region, but they are related to the variables *ankle* and *biceps*, which have low inclusion frequencies (see Figure 1). Therefore, the effect of the observations related to these points are not really interesting from a model building point of view, as *ankle* and *biceps* are not included in any case into the final model. ~~Loosing the visual impression of the single variables importance for the final model may be considered a drawback of this graphic~~Note that the lack of information on the strengths of the variables is a drawback of a plot based on standardized I-frequencies.

4.1.3. O-frequencies and multiple presences

As stated in Section 3.3.1, from the inclusion matrix we can compute, for each observation, the inclusion frequencies based on pseudo-samples which contain the specific observation (I-frequencies) or based on those which do not contain it. The choice of using the former was arbitrary, and potentially one can prefer the use of the latter (let us call them O-frequencies, where "O" stands for "out"). We report in the Web Appendix (Figures A.1 and A.2) the same graphics of Figures 1 and 2 when the O-frequencies are used

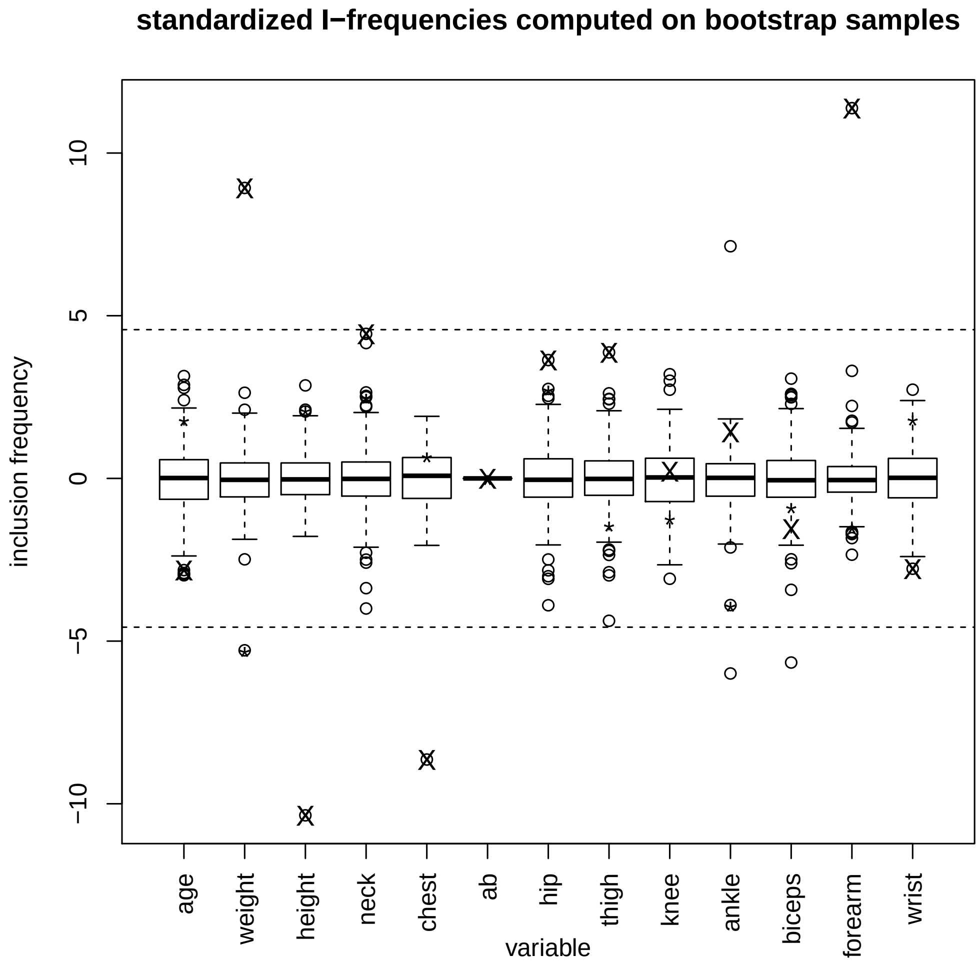


Figure 2: Body fat data: boxplots that summarize the standardized I-frequencies (computed on bootstrap pseudo-samples) obtained for the 13 variables. The symbols "X" and "*" denote the I-frequencies obtained in pseudo-samples (here generated by using bootstrap) which include observation 39 and 221, respectively. The dashed lines delimit the 99% rejection region of a Grubbs' test (including a correction for the multiplicity of the tests).

instead of the I-frequencies. We note that, in this specific example, we obtain similar results. By using the O-frequencies, one should only remember that a smaller O-frequency for an observation x_i means an increasing effect on the inclusion of a variable. Graphically, a lower point means an higher inclusion frequency. This seems counterintuitive and may generate confusion. For this reason, we preferred to use I-frequencies.

The O-frequencies, however, may prove useful in a different analysis. As stated in Section 3.4, when we use bootstrap as a resampling technique, we can separate the pseudo-samples in which a specific observation is included only one time from those in which it is included two or more times. We can then compute two separate inclusion frequencies, that we call I-frequencies-1 and I-frequencies-M, respectively. Together with the O-frequencies (i.e., frequencies computed on bootstrap samples in which the specific observation is included 0 times), these inclusion frequencies can provide us with additional information on the effect of the observation on the inclusion or exclusion of a variable from the statistical model. Pseudo-samples containing two or more times the specific observation are grouped together due to the relatively small amount of cases in which a single observation is repeated three or more times in a bootstrap sample.

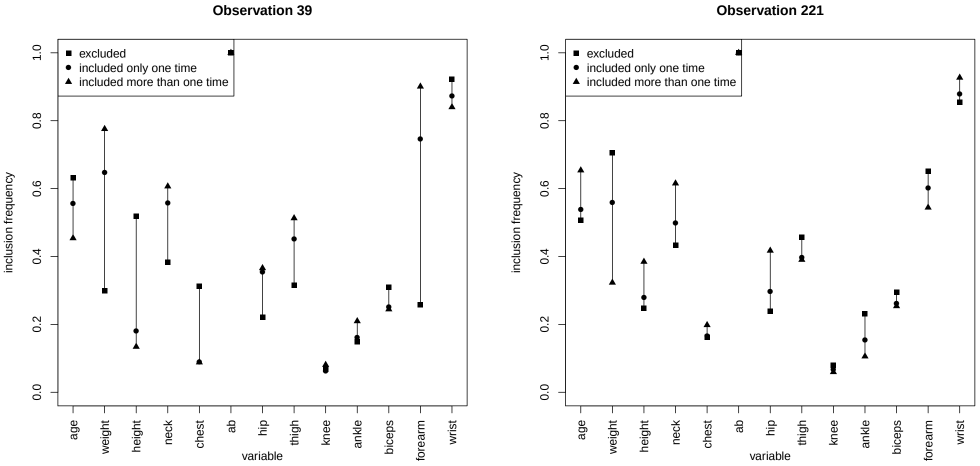


Figure 3: Body fat data: difference in the inclusion frequencies of the 13 variables when observation 39 (left graphic) or observation 221 (right graphic) are excluded (square), included one time (circle) or included more than once (triangle) in the pseudo-samples.

Figure 3 presents the aforementioned frequencies for the 2 possible influential points detected in the previous analyses, namely observations 39 and 221. We immediately notice the different strengths of their influence. Let us focus on the variables *weight* and *height*. The simple presence of observation

39 drastically changes the inclusion frequencies of the two variables, immediately flipping their ranks (when observation 39 is in the sample, *weight* has a higher inclusion frequency than *height*, when it is out of the sample, it is the other way around). A multiple presence of this observation in the sample does not really change the situation, and the differences between I-frequencies-1 and I-frequencies-M are minimal in comparison to the aforementioned differences between I-frequencies-1 and O-frequencies. On the contrary, observation 221 seems to have a much smaller effect. When it is included in the sample, the inclusion frequencies do not change so much (especially for *height*). To notice a certain effect, we should consider the I-frequencies-M. Only when observation 221 is included more than once in the bootstrap sample, indeed, the effect on the inclusion frequencies of *weight* and *height* is strong enough to flip their ranks (and, consequently, their relative chance to be included in the final model).

Note that this procedure allows us to turn into an advantage a possible pitfall of the bootstrap approach. The presence of duplicated observations, indeed, is usually considered a drawback, as pseudo-samples containing duplicated influential points may diverge from the original sample. Robust versions of the bootstrap procedure have been proposed to tackle this issue (see, e.g., Willems & Van Aelst, 2005). Here, instead, we take advantage of this characteristic of the bootstrap to highlight the effect of the possible influential points and to evaluate the strength of their effects. As one may argue that a fair comparison should involve only pseudo-samples without and with only one replication of a specific observation, in the following section we repeat the analyses using subsampling instead of bootrapping as a resampling technique. Pseudo-samples generated by subsampling, indeed, do not contain duplicated observations, as they are drawn without replacement.

4.1.4. Subsampling

As stated in Section 3.4, any resampling technique can be used to generate the pseudo-samples and consequently to compute the inclusion matrix. Figures A.3, A.4 and A.5 in the Web Appendix show the I-frequencies when using subsampling. As long as the size of the pseudo-samples (subsamples) is not extremely high, in this specific example we do not see any noticeable differences with the bootstrap approach (Figures A.3 and A.4). When the subsample size is too large, instead, the differences among samples got lost. Figure A.5, in particular, shows this situation in the case of subsamples of size $n - 2$. When choosing $n - 2$ as a subsample size, it may be more useful to use the O-frequencies instead of the I-frequencies in the analyses. When this is done, the effects of the single observations are taken to their extremes (see Figure A.6 in the Web Appendix). For example, in this dataset the O-frequencies for the variable *weight* are all equal or close to 1, but that related to observation 39, which is 0. For *height*, it is the other way around. On the one hand, having extreme differences among O-frequencies may help to graphically identify the possible influential point in an easier way; on the other hand, several points are equal to 0 or 1, and it is not possible to consider a Gaussian approximation for the distribution of the frequencies. As a consequence, the Grubbs' test cannot be applied. Note that the concept of O-frequencies computed in subsamples of size $n - 2$ is very similar to the idea of delete-2 jackknife, that is sometimes used in the literature related for outlier detection (see, e.g., Martin et al., 2010).

4.1.5. Comparison with FSDA

Figure 4 reports two diagnostic plots obtained through the forward search approach to detect the effect of influential points on model selection by Atkinson & Riani (2002). Although this procedure mainly focuses on the effect of specific observations from a model point of view, rather than a variable point of view, it can be seen as an alternative to our approach, as suggested by a referee. The idea is rather simple: a model is first fitted on a carefully selected subsample of observations, and then stepwise re-fitted on the same subsamples enlarged, at each step, with one of the observation initially excluded (that closer to the fitted model). For details regarding the choice of the initial subsample and the ordering of the observations' inclusion we refer to the original paper. We just point out that the procedure is ~~build~~ built so that the last observations are those with the highest probability to be influential points, being the farthest from the fitted model. Note that the method is robust against the masking effect (see Section 5 for more details), as weaker influential points are considered before the stronger ones.

The left plot of Figure 4, called "deletion statistics plot", shows the variables' importance depending on the subsamples. Each curve visualizes the evolution of the ~~p-value related to a~~ t-test for the nullity of a specific variable's regression coefficient when increasing the number of observations (the farthest observation from the fitted model is the last to be added). Concerning the strongest variables and the most influential observations the results are similar to those obtained with our approach and reported in Figure 1. The variable *ab* plays a predominant role, followed by *wrist*. Moreover, in

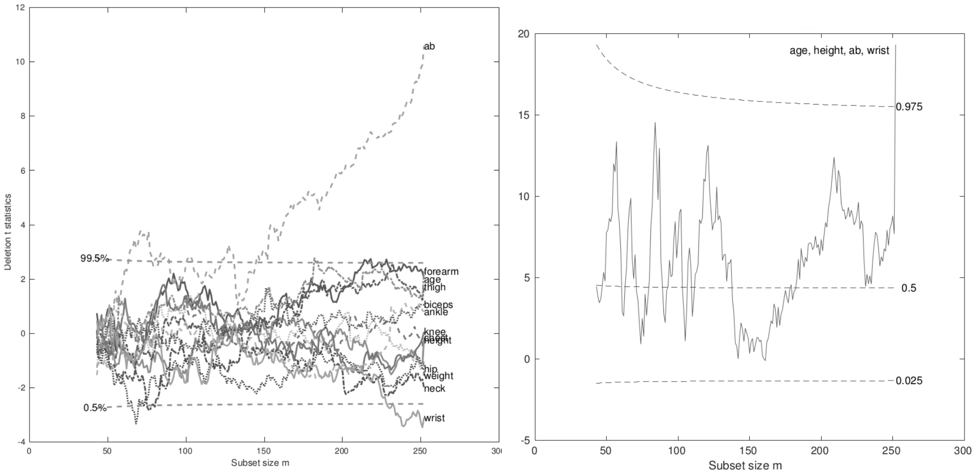


Figure 4: Body fat data: plots from function **FSRaddt** contained inside the MATLAB toolbox FSDA~~routine~~. Left: deletion statistics plot based on the full model; right: Mallow's C trajectory for the model including *age*, *height*, *ab* and *wrist*.

the last step of ~~FSDA~~the forward search, i.e. when the farthest observation, namely observation 39, enters in the sample, the curve related to *forearm* grows drastically, in accordance with the results reported in Figures 1 and 2 (and Table 1). Note that, in contrast to our approach, the deletion statistics plot is based on the full model, which is the starting point of our procedure but which is not selected in any replication.

The right plot of Figure 4, instead, focuses on the model which only includes *age*, *height*, *ab* and *wrist*, i.e. the model selected by backward elimination when observation 39 is excluded from the sample (see Table 1). In particular, the Mallow's C trajectory when increasing the sample size as described above is reported. As expected, this plot shows that the model describes very well the data until observation 39 is included. The Mallow's C trajectory, indeed, is within the 95% bands (dashed lines) until the very last step (i.e., inclusion of observation 39). This result is in line with ours, as we showed that, when observation 39 is in the sample, the selected model is not longer that including *age*, *height*, *ab* and *forearm* but that which includes *weight*, *ab*, *forearm* and *wrist* (see Table 1).

Summing up, in this example FSDA and our approach provide consistent results, as both identify observation 39 as a possible influential point. As mentioned above, however, the focus of the two approaches is relatively different: while the former considers the effect of a specific observation on the goodness of fit of the whole model, our approach focuses on the effect of possible influential points on the inclusion/exclusion of the single variables.

4.2. Myeloma data

In this section we investigate the second dataset. We report only the results obtained when using bootstrap as a resampling technique. As we can see in Figure 5, it seems that there are several points separated from the others. This is a consequence of the small sample size (there are only 48 events). In this situation, especially in a survival context, each observation noticeably influences the model building procedure. To check if these points are influential points, it may be better to rely on the standardized I-frequencies and to compare them with the rejection region of the Grubbs' test.

We report the results in Figure 6. When we consider the rejection region of the Grubbs' test, it is clear that there are no strong influential points in this dataset. The only point that is inside the rejection region and relatively far from the boundary is the largest I-frequency for variable *protein*. This I-frequency is that computed on only pseudo-samples including observation 44. Please note that another point related to this observation is inside the rejection region, namely, the smallest I-frequency for the variable *hemoglob*.

As for the other dataset, we can deepen the analysis on the effect of this observation by comparing the inclusion frequencies computed in pseudo-samples without it (O-frequencies), in pseudo-samples in which it appears only once (I-frequencies-1) and in pseudo-samples in which it appears more than one time (I-frequencies-M). The results are shown in Figure 7. As expected, the most relevant effect is related to variable *protein*. The inclusion of observation 44 in the samples increases the times in which this variable is included in the model. This effect is stronger when the observation is included more than once. In contrast, the presence of observation 44 decreases the inclusion frequency of the variable *hemoglob*. In this case, there is no strong difference whether the observation is included once or more than one time.

As we noted above, in this dataset the effective sample size is quite small and every non-censored observation may have a relatively strong influence on the model-building process. Observation 44 stands out as the most influential one, but there is no strong evidence (e.g., large distance from the 99% rejection region boundaries of the Grubbs' test) that suggests us that it is an outlier. Its presence in the rejection region for variables *hemoglob* and *protein* may simply be related to the type I error of the Grubbs' test.

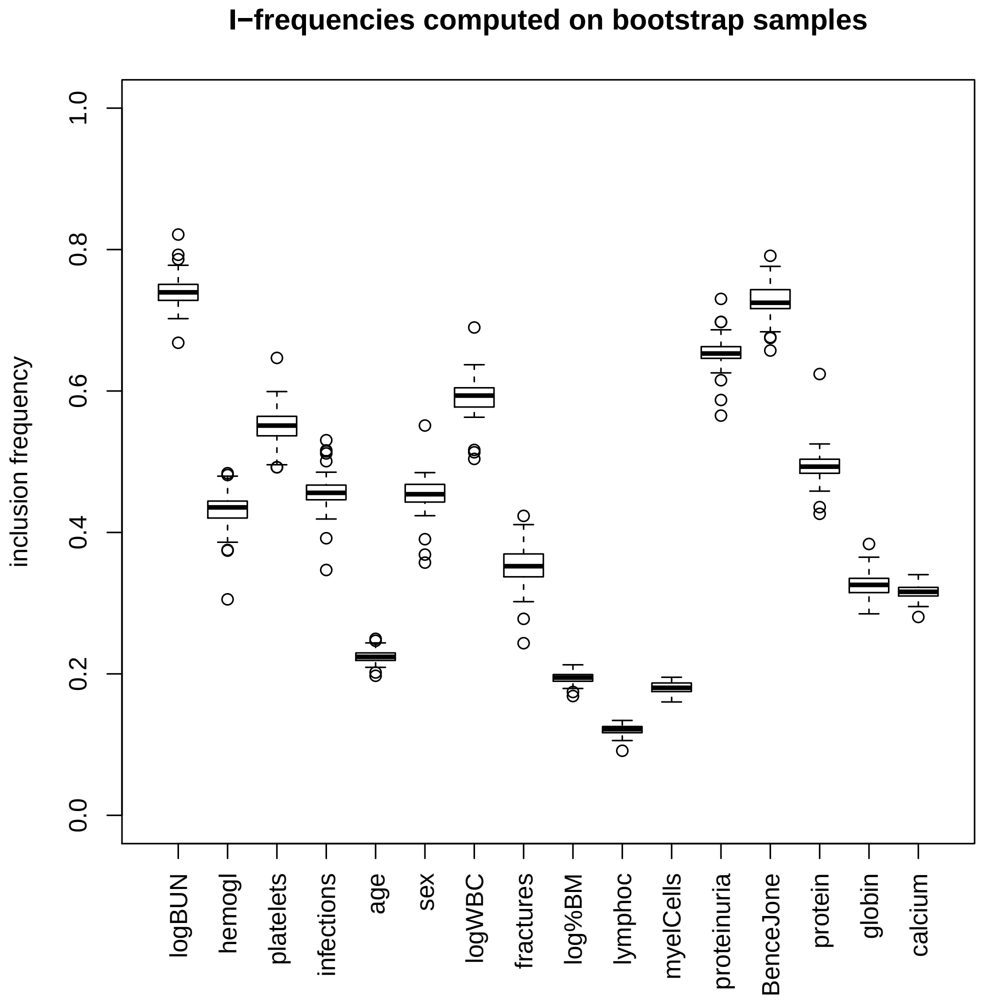


Figure 5: Myeloma data: boxplots that summarize the I-frequencies obtained for the 16 variables, based on boostrap pseudo-samples.

The small sample size also influenced our choice of the bootstrap technique. In order to include an acceptable number of events (here, in particular, the same amount of the original sample) in all pseudo-samples, we implemented a stratified bootstrap, i.e., we resampled separately censored and non-censored observations.

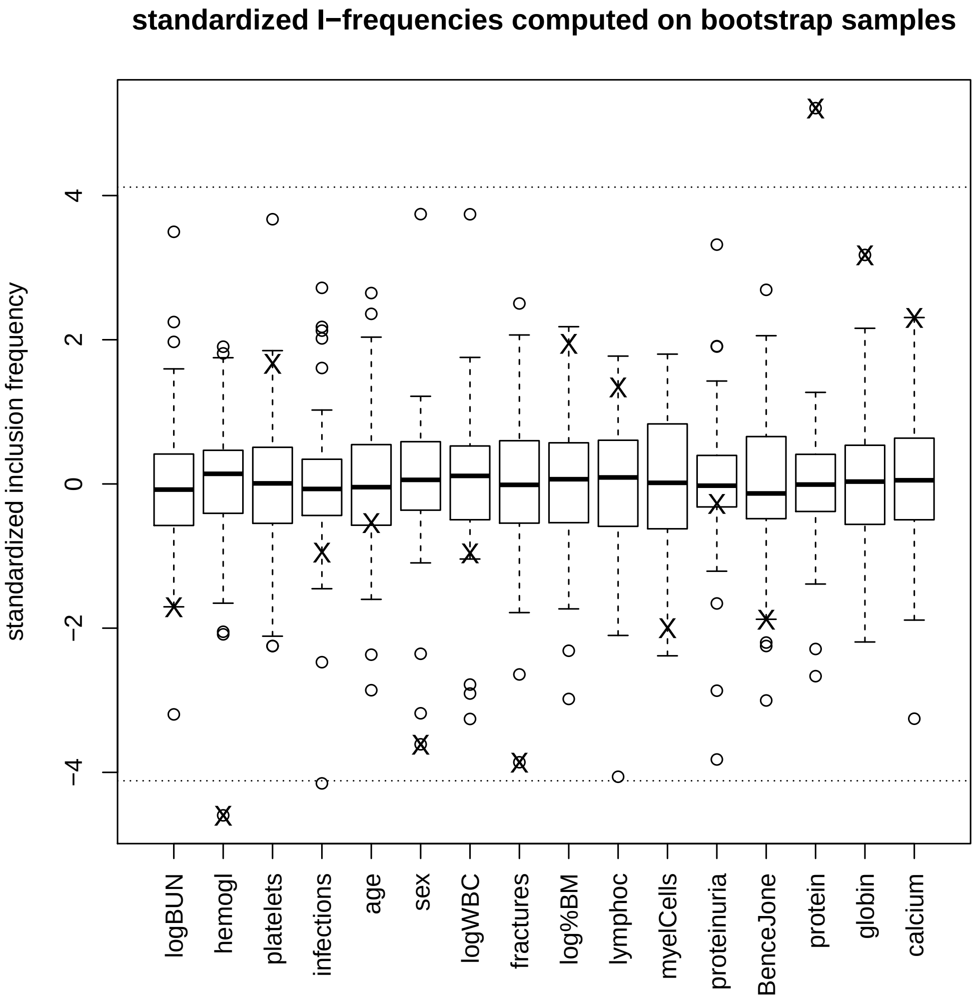


Figure 6: Myeloma data: boxplots that summarize the standardized I-frequencies (computed on bootstrap pseudo-samples) obtained for the 16 variables. The symbol "X" denotes the observed values of the test statistics for observation 44. The dashed lines delimit the 99% rejection region of a Grubbs' test (including a correction for the multiplicity of the tests).

5. Discussion

In this paper we showed how the information present in the inclusion matrix can be used to identify possible influential points. Provided that a resampling based approach is used to select a model, to assess model stability

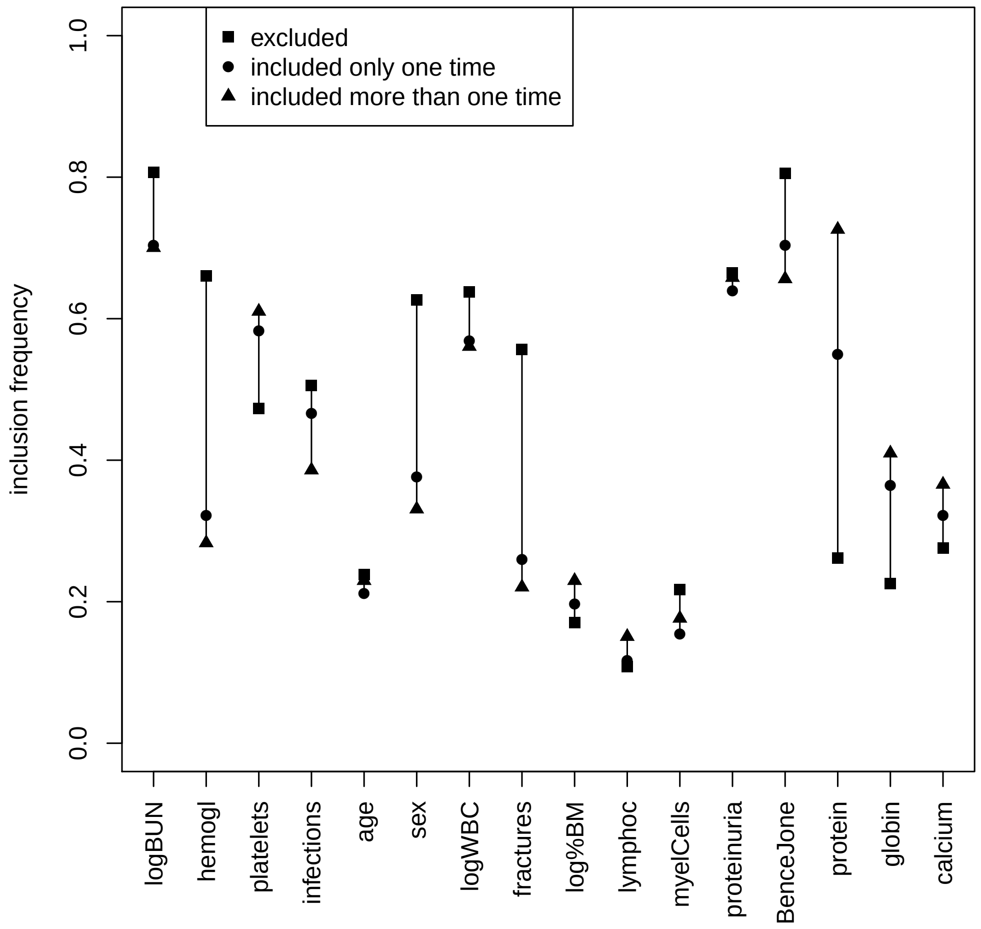


Figure 7: Myeloma data: difference in the inclusion frequencies of the 16 variables when observation 44 is excluded (square), included one time (circle) or included more than once (triangle) in the pseudo-samples.

or to use model averaging for the derivation of a predictor, the generation of the inclusion matrix is a step not requiring any further computation. It uses available information.

Another advantage of our approach is the possibility of having a clear graphical visualization of the results, which allows the user to easily spot possible influential points. As stated in the literature, the identification of an influential point should be done by the user, and not fully delegated to an automatic procedure (Billor et al., 2000).

We considered graphical inspections based both on the simple I-frequencies and on their standardized version. As mentioned before, both approaches have advantages and disadvantages. In particular, when using the simple I-frequencies, we can also have an impression, in the same graphics, of the importance of the variables whose inclusion frequencies may be influenced by a specific observation. As we saw in the first example, this allows us to focus on points of interest, namely influential points which change the selection of the final model, and avoid the investigation of observations that influence the inclusion frequencies of variables which will not be included anyhow. Moreover, the plot of the simple I-frequencies also gives an idea on the variance of their inclusion frequencies, which may be useful in the model building process.

On the other hand, the use of standardized I-frequencies allows us to have a better insight into the influence of the single points. The values of the standardized I-frequencies, indeed, can be contrasted to the rejection region of an univariate test for outliers, in our paper we use the Grubbs' test, to have a more objective estimates of their influence. In studies with very small sample sizes (see the myeloma data) we may identify several observations which seem to be critical. A final assessment needs to consider further criteria from the study.

In this paper we did not consider in detail the problem of the masking effect (Bendre & Kale, 1985). The presence of a strong influential point, indeed, may hide the effect of an observation that has a smaller but still significant influence in the opposite direction. While there are methods, as that by Atkinson & Riani (2002) shown in Section 4.1.5, constructed with this in mind, ours is not specifically designed to tackle the masking effect issue. Nevertheless, we note that in the body fat data example our method was able to detect observation 221 as a possible influential point despite the fact that its effect is the opposite to that of observation 39. Observation 221 has never been identified as an influential point, probably because, in the previous studies, methods that do not take into account the masking effect have been implemented. Our methods, instead, seems to be less prone to suffer from this specific issue.

Acknowledgments

We thank Sarah Tegenfeldt for language improvements and the two anonymous referees for their comments, which led to an improved version of the paper. DBR was partially supported by grant BO3139/4-2 to ALB and WS was supported by grant SA580/8-2. Both grants are from the German Research Foundation (DFG).

Bibliography

Atkinson, A. C., & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer, New York.

Atkinson, A. C., & Riani, M. (2002). Forward search added-variable t-tests and the effect of masked outliers on model selection. *Biometrika*, (pp. 939–946).

Augustin, N., Sauerbrei, W., & Schumacher, M. (2005). The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling*, *5*, 95–118.

Bendre, S., & Kale, B. (1985). Masking effect on tests for outliers in exponential models. *Journal of the American Statistical Association*, *80*, 1020–1025.

Billor, N., Hadi, A. S., & Velleman, P. F. (2000). Bacon: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, *34*, 279–298.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, *53*, 603–618.

Chen, C.-H., & George, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine*, *4*, 39–46.

Chen, C.-H., & Wang, P. C. (1991). Diagnostic plots in Cox's regression model. *Biometrics*, *47*, 841–850.

De Bin, R., Janitza, S., Sauerbrei, W., & Boulesteix, A.-L. (2016). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometics*, *72*, 272–280.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, *21*, 488–506.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, *23*, 881–890.

Gong, G. (1982). Some ideas on using the bootstrap in assessing model variability. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface* (pp. 169–173). Springer New York.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, *21*, 27–58.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*, 1175–1189.

Hansen, B. E., & Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, *167*, 38–46.

Hartigan, J. A. (1969). Using subsample values as typical values. *Journal of the American Statistical Association*, *64*, 1303–1317.

Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*, 879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, *14*, 382–401.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, *4*, 265–266.

Krall, J. M., Uthoff, V. A., & Harley, J. B. (1975). A step-up procedure for selecting variables associated with survival. *Biometrics*, *31*, 49–57.

Kuk, A. Y. (1984). All subsets regression in a proportional hazards model. *Biometrika*, *71*, 587–592.

Martin, M. A., Roberts, S., & Zheng, L. (2010). Delete-2 and delete-3 jackknife procedures for unmasking in regression. *Australian & New Zealand Journal of Statistics*, *52*, 45–60.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*, 417–473.

Myint, P. K., Kwok, C. S., Luben, R. N., Wareham, N. J., & Khaw, K.-T. (2014). Body fat percentage, body mass index and waist-to-hip ratio as predictors of mortality and cardiovascular disease. *Heart*, *100*, 1613–1619.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `https://www.R-project.org/`.

Riani, M., Perrotta, D., & Torti, F. (2012). FSDA: a MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, *116*, 17–32.

Rospleszcz, S., Janitza, S., & Boulesteix, A.-L. (2016). Categorical variables with many categories are preferentially selected in bootstrap-based model selection procedures for multivariable regression models. *Biometrical Journal*, *58*, 652–673.

Royston, P., & Sauerbrei, W. (2007). Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. *Computational Statistics & Data Analysis*, *51*, 4240–4253.

Royston, P., & Sauerbrei, W. (2008). *Multivariable Model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*. Wiley, Chichester.

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L., & Binder, H. (2015). On stability issues in deriving multivariable regression models. *Biometrical Journal*, *57*, 531–555.

Sauerbrei, W., & Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine*, *11*, 2093–2109.

Srivastava, M. (1980). Effect of equicorrelation in detecting a spurious observation. *Canadian Journal of Statistics*, *8*, 249–251.

Su, X., & Tsai, C.-L. (2011). Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*, 261–268.

Tsao, M., & Ling, X. (2012). Subsampling method for robust estimation of regression models. *Open Journal of Statistics*, *2*, 281.

Wan, A. T., Zhang, X., & Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, *156*, 277–283.

Wang, H., Zhang, X., & Zou, G. (2009). Frequentist model averaging estimation: a review. *Journal of Systems Science and Complexity*, *22*, 732–748.

Willems, G., & Van Aelst, S. (2005). Fast and robust bootstrap for lts. *Computational Statistics & Data Analysis*, *48*, 703–715.

Figure 1

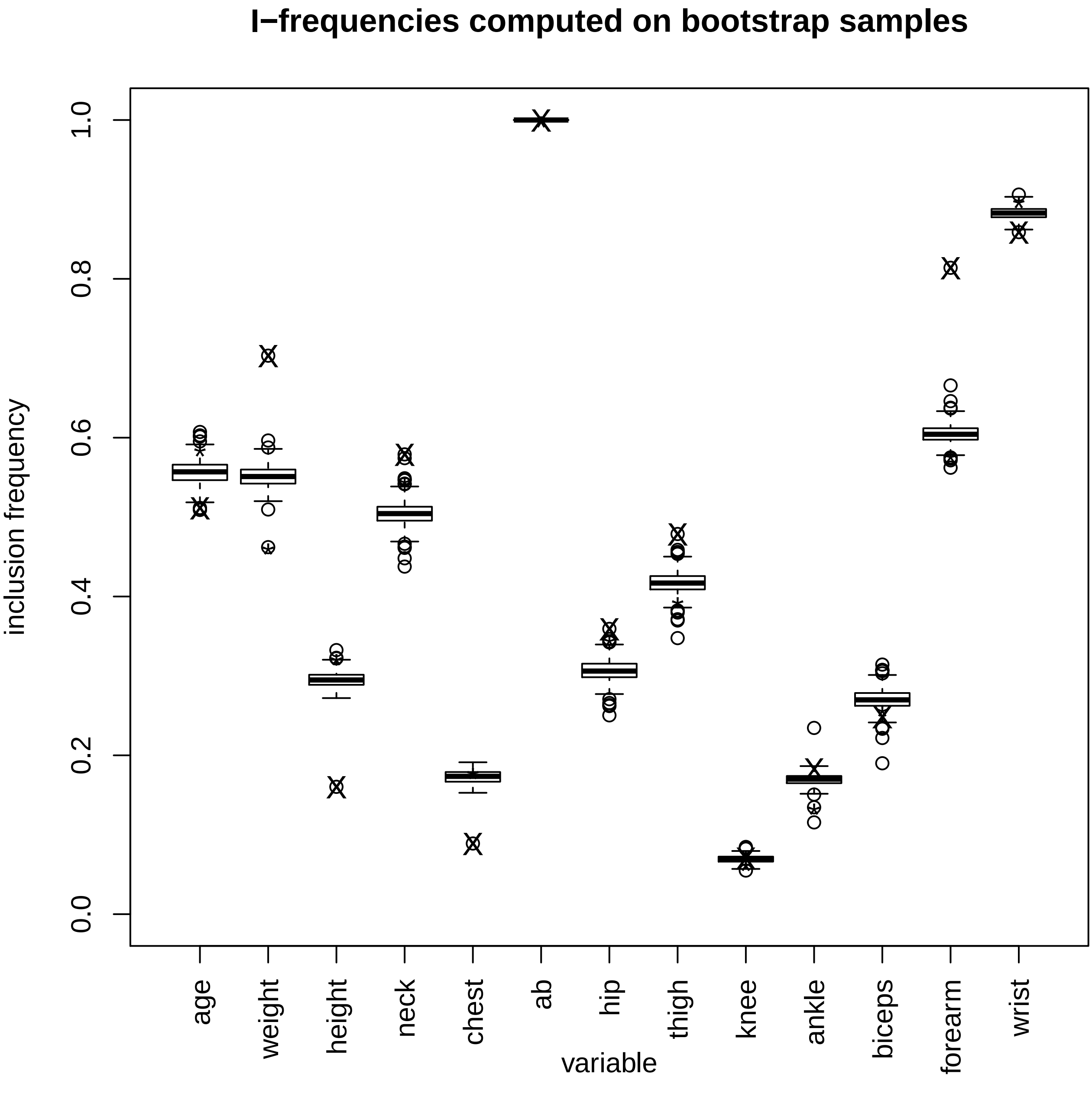

I−frequencies computed on bootstrap samples
inclusion frequency
0.0
0.2
0.4
0.6
0.8
1.0
variable
age
weight
height
neck
chest
ab
hip
thigh
knee
ankle
biceps
forearm
wrist

Figure 2

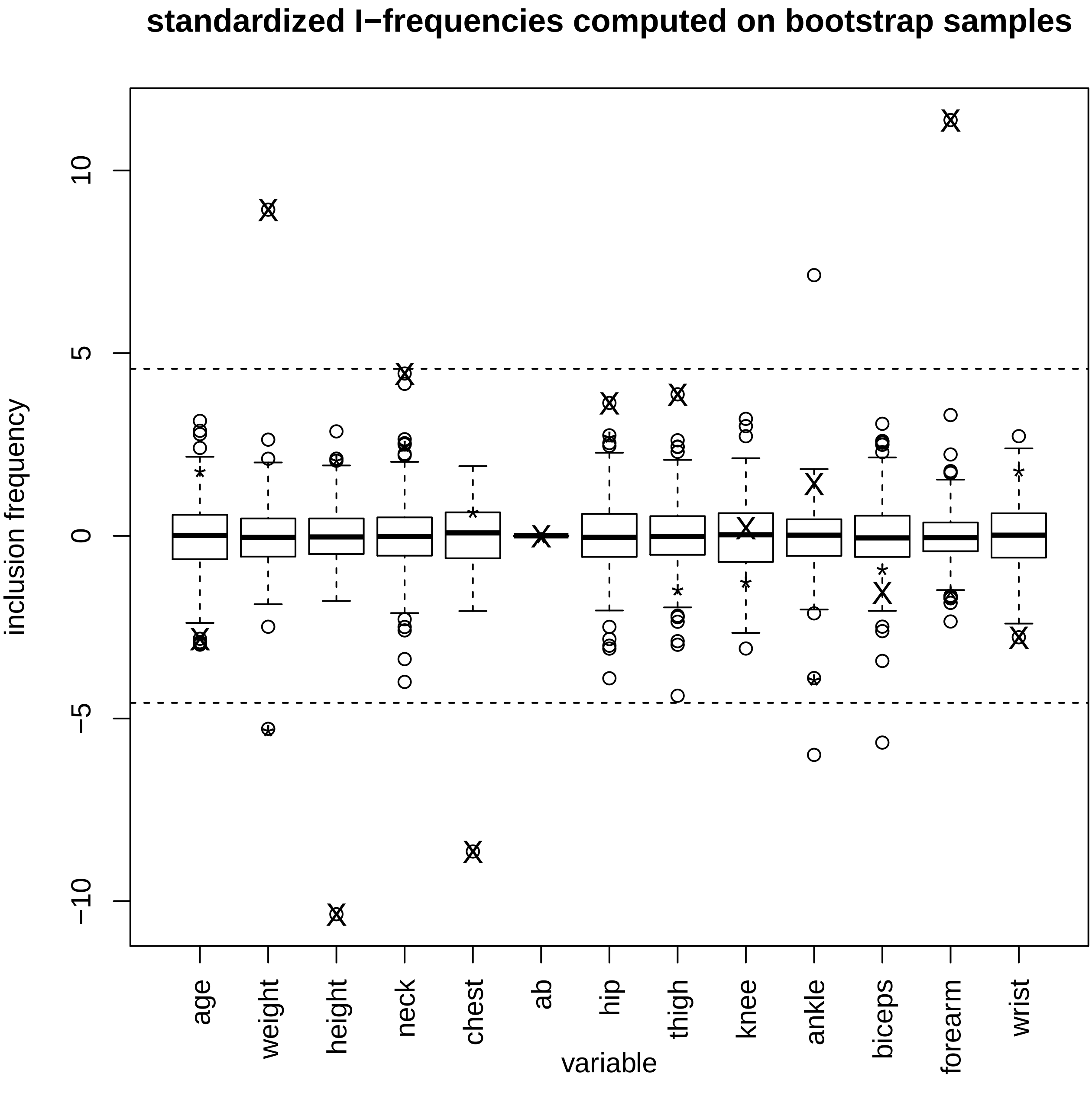
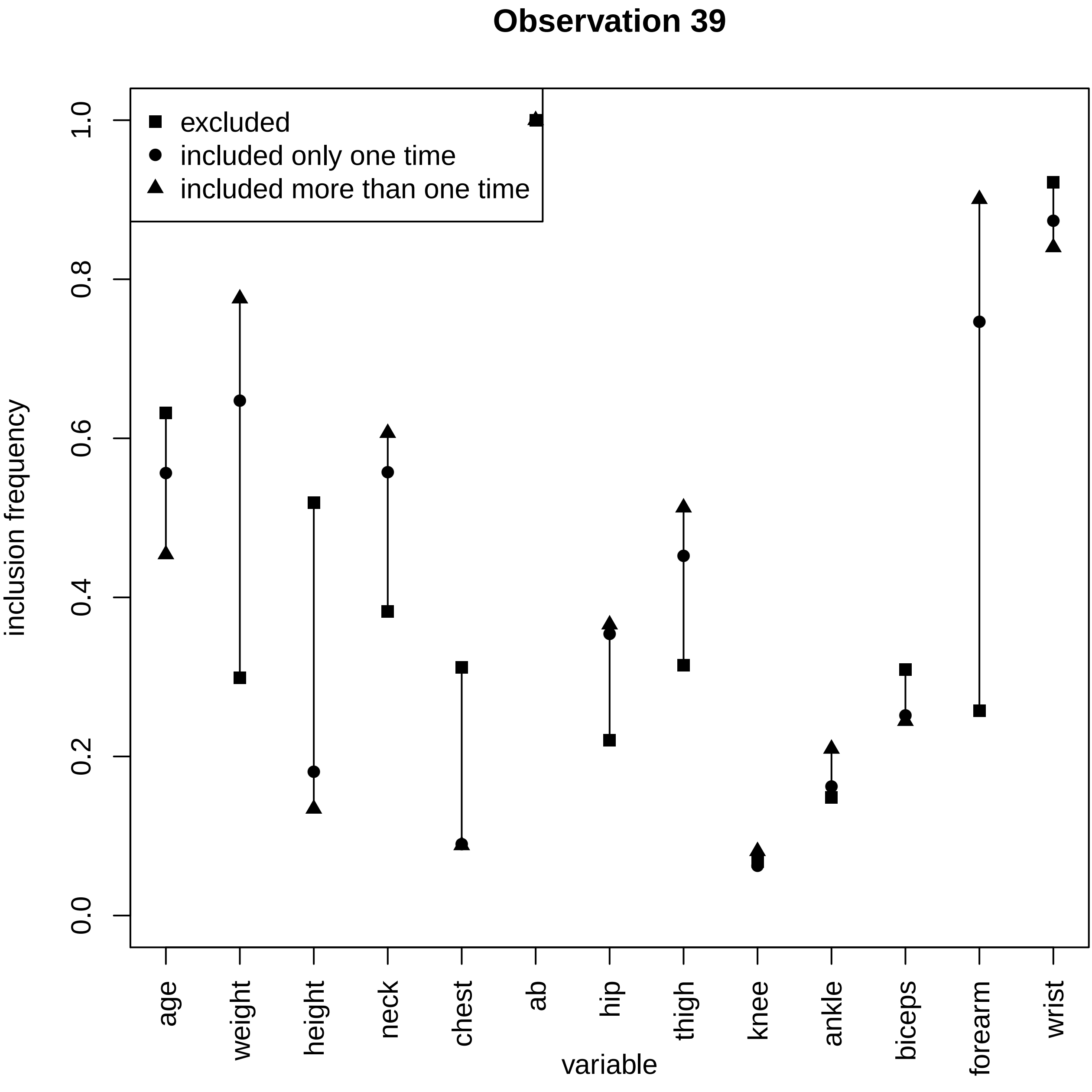

standardized I−frequencies computed on bootstrap samples
inclusion frequency
10
5
0
−5
−10
age
weight
height
neck
chest
ab
hip
thigh
knee
ankle
biceps
forearm
wrist
variable

Figure 3a



Observation 39
■ excluded
● included only one time
▲ included more than one time
inclusion frequency
0.0
0.2
0.4
0.6
0.8
1.0
variable
age
weight
height
neck
chest
ab
hip
thigh
knee
ankle
biceps
forearm
wrist

Figure 3b

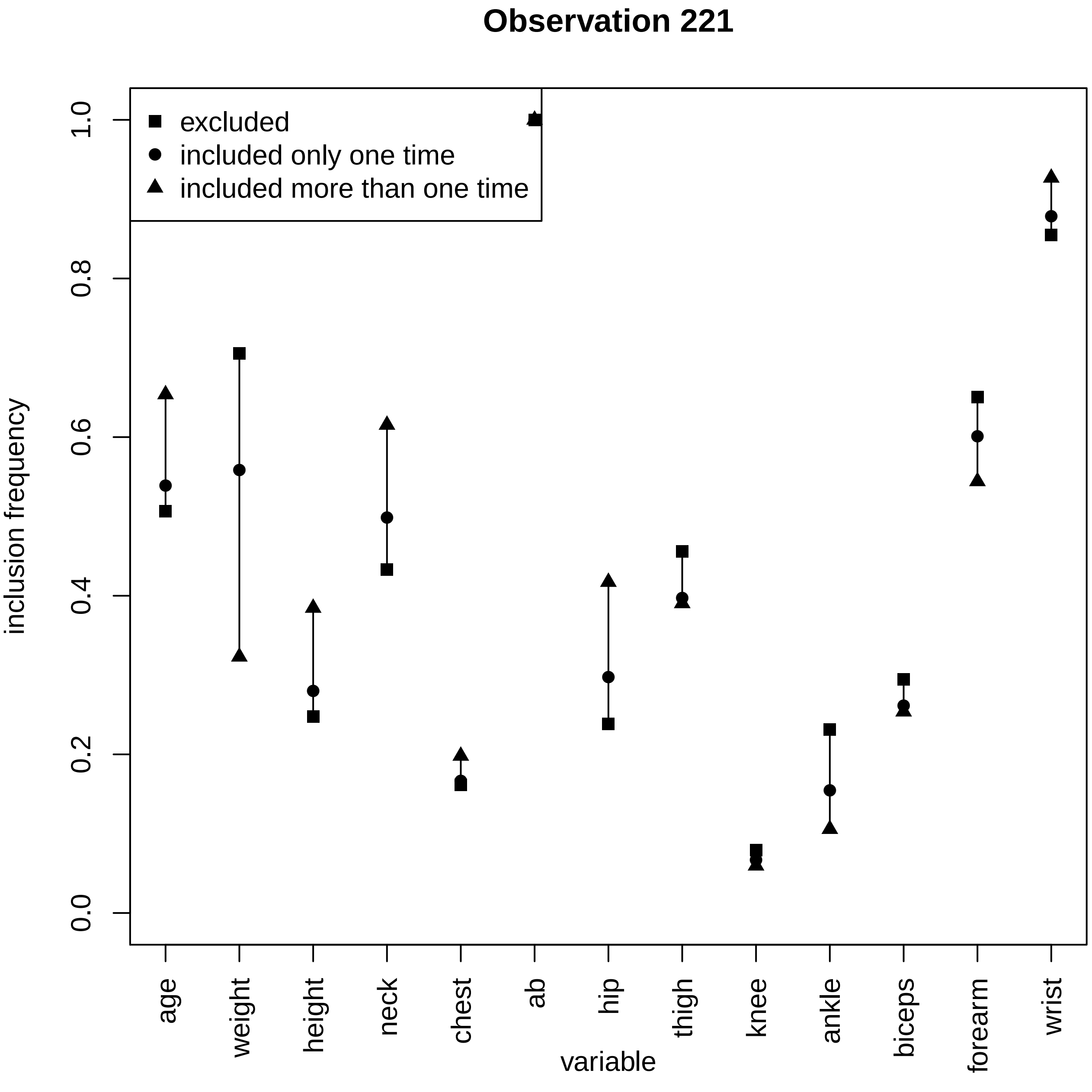
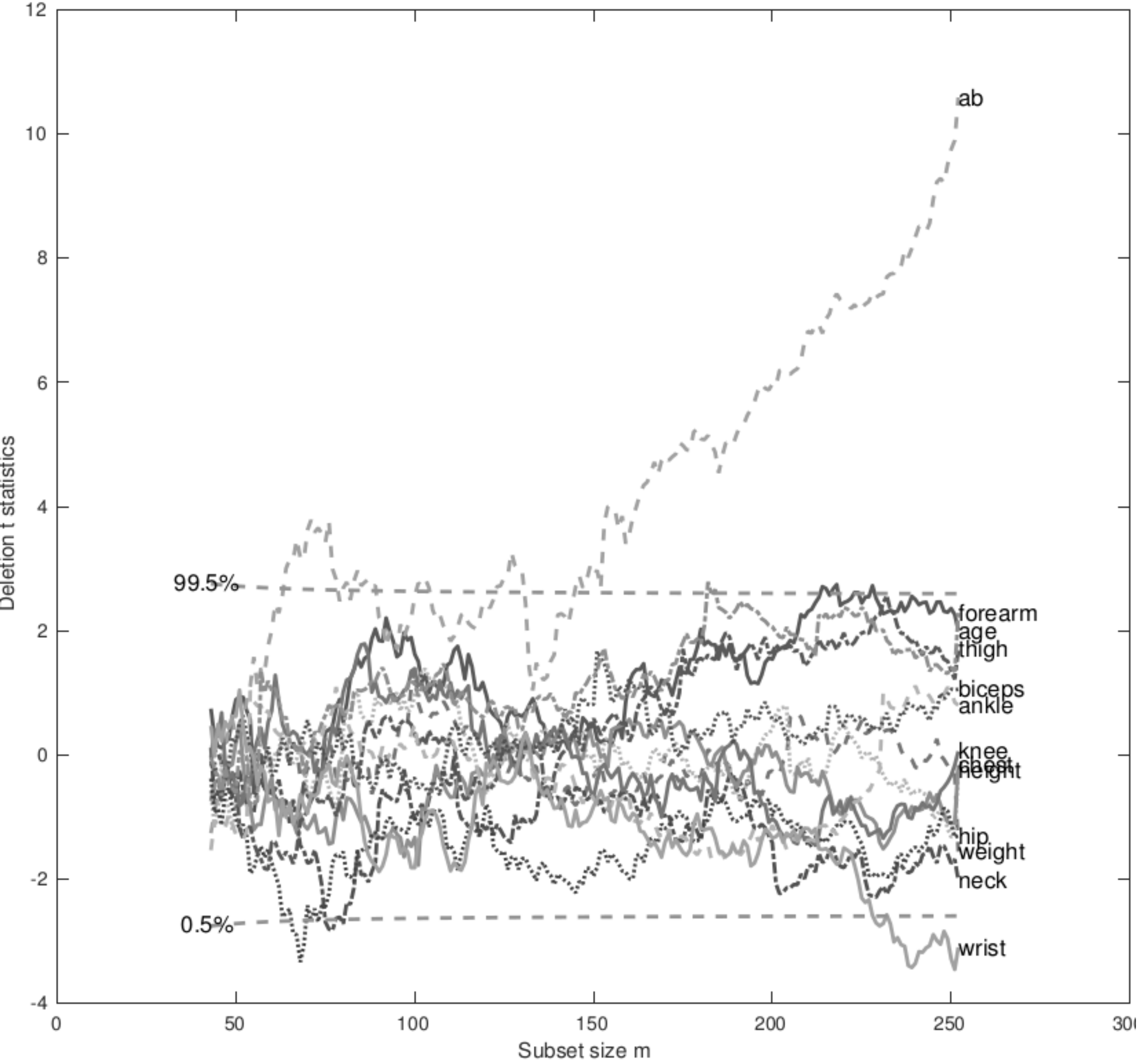

Observation 221
■ excluded
● included only one time
▲ included more than one time
inclusion frequency
0.0
0.2
0.4
0.6
0.8
1.0
variable
age
weight
height
neck
chest
ab
hip
thigh
knee
ankle
biceps
forearm
wrist

Figure 4a



12
10
8
6
4
2
0
-2
-4
Deletion t statistics
0
50
100
150
200
250
300
Subset size m
99.5%
0.5%
ab
forearm
age
thigh
biceps
ankle
knee
chest
height
hip
weight
neck
wrist

Figure 4b

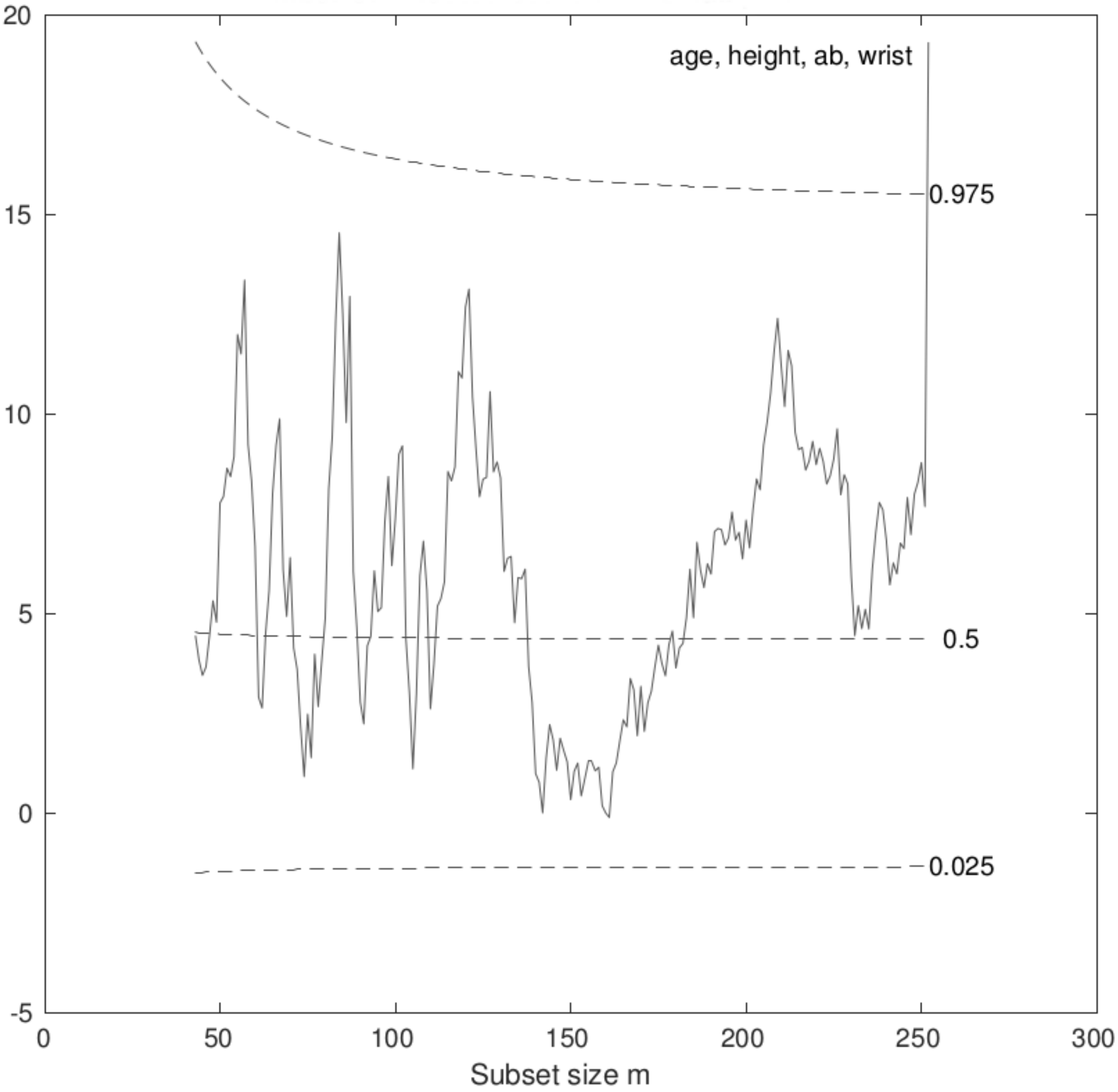

20
15
10
5
0
-5
0
50
100
150
200
250
300
Subset size m
age, height, ab, wrist
0.975
0.5
0.025

Figure 5

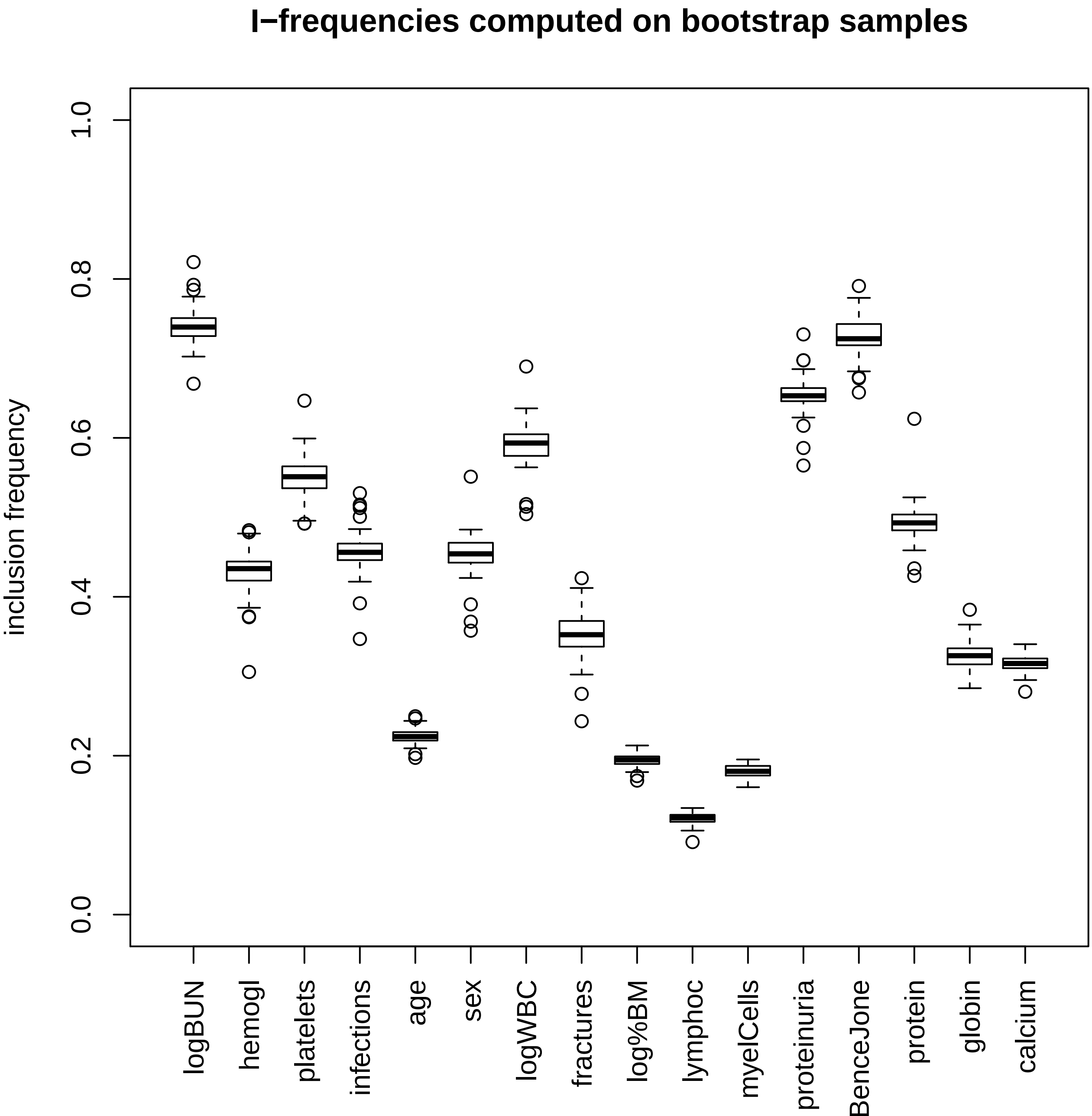
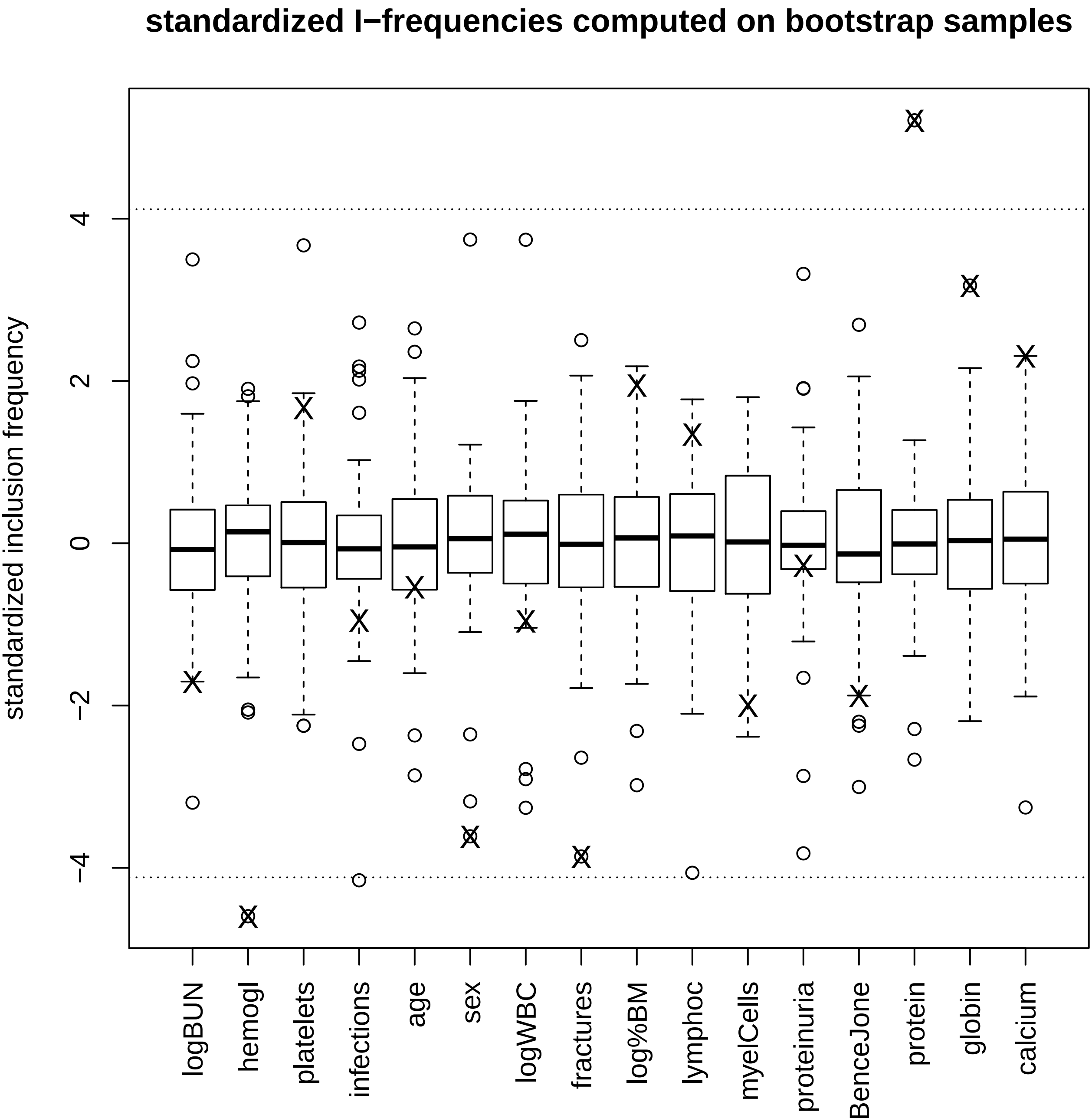
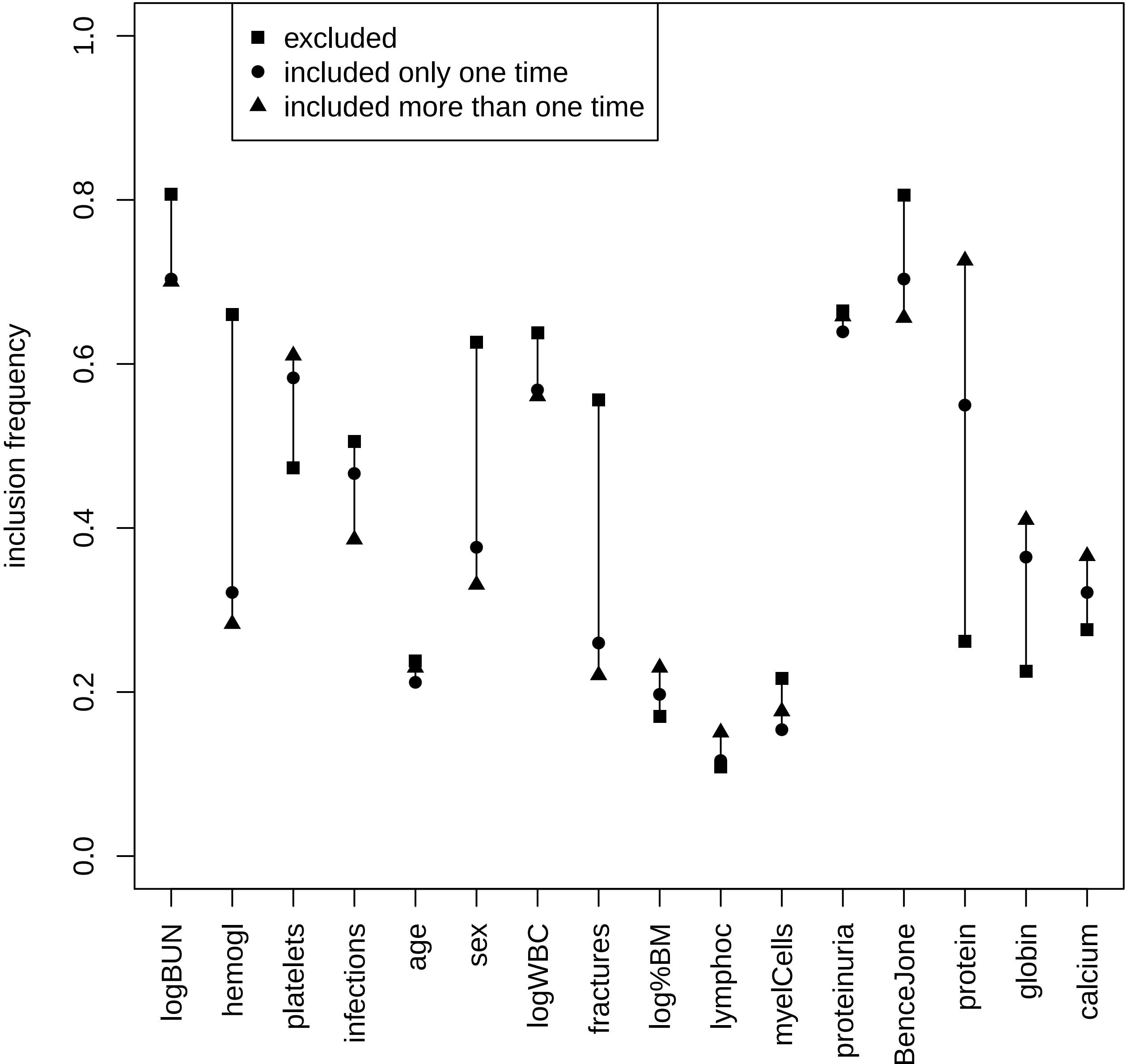

I−frequencies computed on bootstrap samples
inclusion frequency
0.0
0.2
0.4
0.6
0.8
1.0
logBUN
hemogl
platelets
infections
age
sex
logWBC
fractures
log%BM
lymphoc
myelCells
proteinuria
BenceJone
protein
globin
calcium

Figure 6



standardized I−frequencies computed on bootstrap samples
standardized inclusion frequency
4
2
0
−2
−4
logBUN
hemogl
platelets
infections
age
sex
logWBC
fractures
log%BM
lymphoc
myelCells
proteinuria
BenceJone
protein
globin
calcium

Figure 7



■ excluded
● included only one time
▲ included more than one time
inclusion frequency
0.0
0.2
0.4
0.6
0.8
1.0
logBUN
hemogl
platelets
infections
age
sex
logWBC
fractures
log%BM
lymphoc
myelCells
proteinuria
BenceJone
protein
globin
calcium

R-code and dataset(s) (.ZIP)

Click here to download R-code and dataset(s) (.ZIP): R_code.zip