Computational Statistics and Data Analysis 54 (2010) 3181–3193

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A diagnostic method for simultaneous feature selection and outlier identification in linear regression

Rajiv S. Menjoge [a,*], Roy E. Welsch [b]

[a] Operations Research Center, M.I.T., Cambridge, 02139, MA, United States
[b] Sloan School of Management, M.I.T., Cambridge, MA, United States

ARTICLE INFO

Article history:
Received 30 January 2009
Received in revised form 12 February 2010
Accepted 13 February 2010
Available online 20 February 2010

Keywords:
Robust statistics
Forward search
Robust feature selection

ABSTRACT

A diagnostic method along the lines of forward search is proposed to simultaneously study the effect of individual observations and features on the inferences made in linear regression. The method operates by appending dummy variables to the data matrix and performing backward selection on the augmented matrix. It outputs sequences of feature–outlier combinations which can be evaluated by plots similar to those of forward search and includes the capacity to incorporate prior knowledge, in order to mitigate issues such as collinearity. It also allows for alternative ways to understand the selection of the final model. The method is evaluated on five data sets and yields promising results.

1. Introduction

Outliers substantially complicate the already difficult task of model selection in linear regression. The question of which features to select as well as how many of the chosen features to select can both be grossly influenced by outliers, and to make things even tougher, the features that are selected in a model will influence which observations are considered outliers.

Robust model selection, however, brings complications even beyond its statistical framework. For instance, one complication of outlier detection is that a point which statistical methods deem an outlier could in fact be the most important observation in the data set depending on the application and the cause for the outlier. Forward search is one remedy for this in that it identifies outlying points of various magnitudes and creates plots that highlight the effect of each observation on various inferences made in linear regression. It thereby provides several possible good models and gives the analyst a way to visualize these. The goal of this paper is to extend these ideas to the case of simultaneous feature selection and outlier detection, which has arisen more recently in the literature.

We organize the rest of the paper as follows: the remainder of this section provides a brief literature review, discussing other ways the problem of simultaneous feature selection and outlier detection has been tackled and how diagnostic methods have evolved. Section 2 provides a review of forward search. Section 3 discusses the method that we propose in this paper, which we call backward selection search. Section 4 presents the output of our method on five well-known data sets. Finally, conclusions are given in Section 5.

1.1. Literature review

In the sections below, we provide a literature review for robust model selection and diagnostic methods for outlier detection. We then discuss how our method fits in with the existing literature.

* Corresponding author.
E-mail addresses: menjoge@mit.edu (R.S. Menjoge), rwelsch@mit.edu (R.E. Welsch).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.02.014

1.1.1. Diagnostic methods for outlier detection

Diagnostic methods for outlier detection aim to separate observations into a "clean" set and a set of possible outliers (Hadi, 1992; Hadi and Simonoff, 1993), as well as to highlight the effects of these possible outliers. Some of the earliest diagnostic methods include leave-one-out deletion techniques, which study the effect of each individual observation on the inference (Belsley et al., 1980) by removing one of the data points and calculating the statistics of interest with and without this point to see if there is a large difference. However, this tends to fail when outliers come in groups. In general, the phenomena of outliers going undetected because of the presence of another set of outliers (Masking) and "good" observations being misidentified as outliers because of the presence of a set of outliers (Swamping) (Hadi, 1992; Hadi and Simonoff, 1993; Atkinson, 1986b; Fung, 1993), make the aim of diagnostic methods difficult to accomplish.

Over time, however, progress has been made and a popular recent algorithm, proposed by Atkinson and Riani (2000), called forward search does a good job of detecting outliers and highlighting the influence of these outliers on various regression statistics. Forward search starts with a small subset of q "good" points, where q is the number of parameters (generally $q = p + 1$ where p is the number of features). Iteratively, the points which adhere most to the pattern that the good points follow are added to the set of "good" points, until all points are in the set of "good" points. The output is a sequence of sets of points of sizes $q, q + 1, \ldots, n$ along with the statistics of interest for these sets. Plots are then made where the y-axis is a statistic of interest and the x-axis is the size of the subset. Details are given in Section 2.

1.1.2. Robust model selection

Some of the earliest papers in simultaneous outlier detection and feature selection focus on developing criteria for an optimal robust model: Ronchetti (1985) and Ronchetti and Staudte (1994) proposed robust versions of the selection criteria AIC and C_p respectively. Meanwhile, Ronchetti et al. (1997) proposed robust model selection by cross-validation.

More recent papers have explored the additional issue of selecting a sequence of relevant features in a robust manner: Khan et al. (2007) proposed a way of replacing various statistics with their robust counterparts in the LARS algorithm to perform robust LARS; Morgenthaler et al. (2004) formulated outlier detection as a variable selection problem on an augmented matrix. McCann (2005), McCann and Welsch (2007), and Kim et al. (2008) have built extensions on this using LARS and best subsets respectively to perform the variable selection. McCann and Welsch (2004) also used the idea of feature selection on an augmented matrix to propose an alternative way to select sequences of points for forward search. In addition, Atkinson and Riani (2002) enhanced their idea of monitoring variable coefficients and significance in the forward search through an added variable t-test for variable selection in the context of regression.

1.1.3. Our approach

We seek to combine the two objectives of diagnosing outliers and selecting features into one method which simultaneously selects one parameter for both feature set size and observation set size, and produces a sequence of reasonable feature-observation models, which can be combined with prior knowledge and visualized for further inspection. Our method is most similar to (Morgenthaler et al., 2004) and its extensions, while our analysis mechanism tries to replicate that of forward search.

2. A description of forward search

Forward search is a diagnostic method that produces plots which help identify outliers, their structure, and their effect on various statistics of interest. As mentioned before, it starts with a small subset of q "clean" points, where q is the number of parameters (generally $q = p + 1$ where p is the number of features). Iteratively, the points which adhere most to the pattern that the "clean" points follow are added to the set of "clean" points, until all points (including outliers) are in the set of "clean" points. The output is a sequence of sets of points of sizes $q, q + 1, \ldots, n$ along with the statistics of interest for these sets. Plots are then made where the y-axis is a statistic of interest and the x-axis is the size of the subset. The following is the procedure in detail:

1. Start by identifying an initial subset of q "clean" points using a high breakdown robust method. Atkinson and Riani suggest using Least Median of Squares in order to find an initial fit and then selecting the q points with the smallest squared residuals with respect to the initial fit to be the q "clean" points. Least Trimmed Squares is a reasonable alternative to Least Median of Squares, but both methods have been found to pick good initial subsets.
2. In a typical iteration, where the "clean" subset contains m points: conduct ordinary least squares using those m observations. Then compute all statistics of interest, say θ_m. Find the squared residuals for all points (not just the ones in the good set) and let the updated "clean" subset contain only the $m + 1$ points with the smallest squared residuals to this fit. It should be noted that these sequences of points are not necessarily nested.
3. Repeat step 2 until all the points are added, so that we are left with a vector $\theta = \{\theta_{q+1}, \theta_{q+2}, \ldots, \theta_n\}'$. The output of this procedure is typically a plot of θ_m as a function of m.

The purpose of the plot is to study how various outliers affect the statistics of interest. As a simple example, Fig. 1 shows a plot of the Hertzsprung–Russell Star data set, that will be discussed further in Section 4. It is an example where leave-one-out deletion techniques fail. Figs. 2 and 3 illustrate the visual output that forward search provides for this data set.