Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2006/08

Robust ANalysis Of VAriance: an approach based on the *Forward Search*

Bruno Bertaccini,
Roberta Varriale

Università degli Studi
di Firenze

Robust ANalysis Of VAriance an approach based on the *Forward Search*

Bruno Bertaccini Roberta Varriale
Department of Statistics "G. Parenti"
University of Florence (Italy)[1]

Abstract

We present a simple robust method for the detection of atypical observations and the analysis of their effect in the ANOVA framework. We propose to use a *forward search* procedure that orders the observations by their closeness to the hypothesized model. The procedure can be applied following two different strategies: one that adds units maintaining the relative group dimension and the other that adds only one new unit at each step of the search. The assessment of the goodness of the method is carried out through a simulation study. The method is then applied to a dataset collected by the Italian National University Evaluation Committee for the evaluation of the effectiveness of the degree program reform applied during the academic year 2001/02. Results are always presented through easy to interpret plots which are powerful in revealing the structure of the data.

Key words: ANOVA, Fisher *F* test, forward search, graphical methods outliers, policy effectiveness, robustness.

1. Introduction

One of the most important topic in statistical inference is the presence of outliers in the data. Outliers can be defined as observations which appear to be inconsistent with the reminder of that set of the data (Hampel *et al*., 1986; Staudte and Sheather, 1990, Barnett and Lewis, 1993). Outliers can be *contaminants*

1 To contact the authors, send an e-mail to: brunob@ds.unifi.it

(arising from other distributions) or can be *atypical* observations generated from the assumed model (see also Barnett, 1988). They often can be masked and should always be examined to see if they follow a particular pattern, come from recording errors, or could be explained adequately by alternative models. Although outliers are often synonymous with "bad" data, they are frequently an important part of the data. They need a very special attention because a small departure from the hypothesized model can have strong negative effects on classical estimators efficiency (Tukey, 1960).

The key statistic in the ANOVA is the F test of group means difference. The test is very powerful under classical assumptions, but it is strongly affected by the presence of outliers, due to the fact that it is based on the sample group mean that is not a robust statistic.

The purpose of this article is to implement the *Forward Search* method in the ANOVA framework. The method, first proposed by Atkinson and Riani (2000) for linear regression models, is a general powerful approach for detecting and investigating the effect of observations that differ from the bulk of the data. The starting point is to fit the model to very few observations chosen in some robust way, order all the observations by their closeness to the fitted model, increase the subset size, refit and continue until all the data are entered in the model. Through the search we try not only to identify the outlying observations but also to analyse their effect on the estimation of parameters and on inferences on the model. Our proposal is thus the following: at every step of the search we compute parameters estimates, residuals, classical F values and all the other considerable statistics. The collection of these information is analysed (graphically or otherwise) in order to identify a *cut-off* point that divides the group of outliers from the other observations. Since at the moment there are no rules that allow the automatic identification of this point; we advocate the use of a graphical approach. The F value obtained at the cut-off point should be used to get a robust F_{FS} test. In this context, the robustness of the method does not derive from the choice of a particular estimator with an high breakdown point but from the progressive inclusion of units into a subset which, in the first steps, is outlier free.

After a brief review of the classical one-way ANOVA, in paragraph 3 we underlay the problems of the classical F test in presence of atypical observations. In the paragraph 4, we present the proposed *Forward Search* algorithm in the

ANOVA framework, showing the application and the advantages of the proposed approach in identifying outliers through a simulation study. In paragraph 5 we illustrate the results of a Montecarlo study simulating a forward search analysing the significance of the F_{FS} test. Finally, we show an application of the proposed approach to real data, using a set of information referring to the performance of the Italian university system.

2. Univariate One-Way Analysis of Variance (a review)

The ANalysis Of VAriance is one of the most widely used statistical techniques to test means difference of several populations.

In this paper we are going to study the simplest type of ANOVA, the *one-way* ANOVA typically characterized by a sample of n_i observations $(i=1..\ g)$ from each of the *g* normal populations with equal variance.

The model for each observation is

$$y_{ij}=\mu+\alpha_i+\varepsilon_{ij}$$
$$=\mu_i+\varepsilon_{ij}, \qquad (j=1,2,...,n_i);$$

where μ is the common mean level of the treatments and $\mu_i=\mu+\alpha_i$ is the *i*-th population mean[2].

Classical assumptions on this model are:

$E(\varepsilon_{ij})=0$

$var(\varepsilon_{ij})=\sigma^2<\infty$ for all i, j,

ε_{ij} are independent and normally distributed.

So, the *g* samples are assumed to be independent. The ANOVA provides a useful way of thinking about the way in which different treatments affect a measured variable – the idea of allocating variation to different sources. This idea can be summarised by the decomposition of the total deviance (DT) in *'within*

2 It is common to add the restriction that $\sum_i^g \mu_i=0$, in order to make the model identifiable.

groups' (DW) and *'between groups'* (DB) deviance. If the hypotheses are true, all y_{ij} come from the same population $N(\mu, \sigma^2)$; so, through *DW* and *DB*, we can obtain two unbiased estimates of σ^2 based the first one on the sample variances $s_1^2, s_2^2, \ldots, s_g^2$, and the second one on the sample means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_g$.

From *DW,* we have that:

$$E(DW)=E\left(\sum_i^g \sum_j^{n_i} (y_{ij}-\bar{y}_i)^2\right)=E\left(\sum_i^g (n_i-1)s_i^2\right)=(n-g)\sigma^2$$

and from *DB* we obtain:

$$E(DB)=E\left(\sum_i^g n_i(\bar{y}_i-\bar{y})^2\right)=E\left(\sum_i^g n_i(\bar{y}_i-\mu+\mu-\bar{y})^2\right)=\ldots=(g-1)\sigma^2 \cdot$$

Under the null hypothesis, the decomposition of the total deviance is a partitioning of a *Chi-square* random variable: when scaled from their degrees of freedom, *DT*, *DW* and *DB* are distributed, respectively, as a $\chi^2_{(n-1)}$, $\chi^2_{(n-g)}$ and $\chi^2_{(g-1)}$. This partitioning is true only if the *DW* and the *DB* are independent, which follows from the normality in the ANOVA assumptions.

The ratio statistics

$$F=\frac{DB/(g-1)}{DW/(n-g)}\sim F_{(g-1),(n-g)}$$

only if the null hypothesis is true. Thus, the ANOVA *F* test is a function of the variance of the set of group means, the overall mean of all the observations, and the variances of the observations in each group weighted for group sample size. The larger the difference in means, the larger the sample sizes, and/or the lower the variances, the more likely a finding of significance.

3. Problems relating to the presence of outliers

Due to the presence of the statistic $\bar{y}_i$ in both the *DW* and *DB*, the value of the *F* statistic is strongly affected by the presence of outliers. The sample mean, under normality assumptions is, in fact, the best unbiased estimator of the population

mean but it shows a strong loss of efficiency in case of contamination or misspecification of the model[3].

Consider, for example, a sampled dataset with 50 observations in each of 3 groups (a *balanced* one-way ANOVA). The units in each group are generated by a Standard Normal distribution, but the second group is heavily contaminated by 10 observations coming from a Uniform distribution $U(10, 11)$. Obviously, in this situation outliers are so different that are easily identifiable by any other approach: this difference is stressed just to illustrate clearly our proposal.

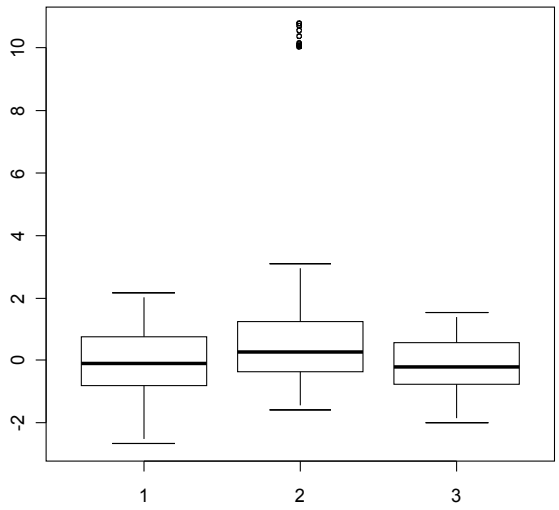

Figure 1. *Boxplot that shows the composition of the generated example: group2 is strongly contaminated.*

Figure 1 shows the results of this sample. Here, the F statistic has 2 and 147 degrees of freedom and its value (12.001 with p-value of $1.48 \cdot 10^{-5}$) falls, due to the contamination, in the rejection region.

Also in the presence of more realistic contaminations, test F will often leads us to erroneously reject the null hypothesis. Let us consider, for example, datasets with three balanced groups of increasing sample size n_i

3 Only one unit can be moved towards infinity to cause an arbitrarily large change in the estimate of μ : the breakdown point of this estimator is zero.

$(n_i = 20, 40, \ldots, 200;\ i = 1,2,3)$ observations coming from a Standard Normal distribution. Only the second group is contaminated by increasing rates ϵ $(\epsilon = 0.05, \ldots, 0.10)$ from a $N(2, 1)$ distribution; so the distribution of this group is $(1-\epsilon)N(0,1)+\epsilon N(2,1)$.

Table 1 shows the relative frequencies $rf_{(n_i,\epsilon)}$ over 10000 simulations in which the *F* test falls in the rejection area at the nominal significance level of $\alpha = 0.05$, giving an approximation of the true *type I* error probability:

$$Pr\left(F_{(n_i,\epsilon)} > F_{0.95;\, df_1=2,\, df_2=3(n_i-1)}\right) \approx rf_{(n_i,\epsilon)}.$$

For example, for the pair $(n_i = 100,\ \epsilon = 0.08)$ there are 8 contaminants in the second group, and the classical test *F* reject the null hypothesis 1629 times, giving an approximate α value of 0.1629.

The larger is n_i and the bigger is ϵ, the stronger is the effect of the contamination on the *F* test.

	$\epsilon =$	**5%**	**6%**	**7%**	**8%**	**9%**	**10%**
$n_i =$	**20**	0.0492	0.0492	0.0492	0.0648	0.0648	0.0648
	40	0.0572	0.0572	0.0733	0.0733	0.1018	0.1018
	60	0.0687	0.0845	0.0845	0.1080	0.1080	0.1441
	80	0.0761	0.0985	0.1240	0.1240	0.1548	0.1921
	100	0.0882	0.1106	0.1348	0.1629	0.1982	0.2389
	120	0.970	0.1177	0.1405	0.2018	0.2386	0.2801
	140	0.1096	0.1293	0.1839	0.2175	0.2926	0.3368
	160	0.1191	0.1651	0.1933	0.2555	0.2934	0.3735
	180	0.1297	0.1731	0.2328	0.2663	0.3394	0.4262
	200	0.1441	0.1895	0.2468	0.3142	0.3923	0.4689

Table 1. *Approximation of the true type I error probability of the F test in presence of contamination.*

4. Forward Search in the Analysis of Variance

One methodology useful not only to detect and investigate observations that differ from the bulk of the data, but also to analyse their effect on the estimation of parameters and on aspects of inference about models is the forward search, proposed by Atkinson and Riani in year 2000. The basic idea of the "forward" procedure is to order the observations by their closeness to the fitted model. The starting point is to fit the model to very few observations chosen in some robust way, order all the observations by their closeness to the fitted model, increase the subset size and refit the model. The process continues with increasing subset sizes until, finally, all the data are fitted.

During the search, at each stage, it is possible to monitor some quantities, such as parameter estimates, residual plots, *F* value and other informative statistics, to guide the researcher in the identification of the outliers. In the absence of outliers, for example, both parameter estimates and residuals are expected to remain sensibly constant during the search; in the presence of outliers, instead, this quantities will remain constant until the outliers enter the subset to be fitted.

The forward search algorithm is made up of three steps: the first concerns the choice of an initial subset, the second refers to the way in which the forward search progresses and the third relates to the monitoring of the statistics during the search.

The methodology used in this paper is adapted to the peculiarity of the model under study, in particular it has to take into consideration the presence of groups in the data structure of the model. We implemented a proportional and non proportional approach: the difference is basically in the number of units that join the model during the search and points out some characteristics of the data structure. Furthermore, we proposed a procedure to obtain a robust forward *F* test, individuating a cut-off point of the collection F_{FS} of the classical *F* test in each step of the search that divides the group of outliers from the other observations. We derived from a Montecarlo simulation study that with the proposed method the probability of the the *type I* error is lower than with the classical ANOVA.

Programming codes for R and S-Plus, developed by the authors, are available on demand.

4.1. *Step 1*: choice of the initial subset

The first step of the forward approach is the choice of an initial outlier free subset of observations. Many robust methods were proposed to sort the data into a clean and a potentially contaminated part; our proposal[4] in the ANOVA framework is to start with the observations y_{ij} that satisfy:

$$min\left|y_{ij}-med_i\right|$$

in each group *i,* $(i=1..g)$, where med_i is the group *i* sample median.

The dimension of the initial subset $S^{(*)}$ is, therefore, almost surely equal to *g*. Since we need at least $g+1$ observations in order to have residual degree of freedom for the estimate of the standard error, the algorithm enters in the starting subset, the next unit with the minimum residual from the medians.

4.2. *Step 2*: adding observation during the search

At each step, the forward search algorithm adds to the subset the observations that are closer to the previously fitted model. This can be accomplished following two different strategies: the first, called *non proportional*, adds just one new unit at each step, while the other, *proportional*, enters the minimum number of observations necessary to respect the overall composition (the groups proportions) of the sample.

Formally, in the first procedure, given the subset $S^{(m)}$ of dimension $m=\sum_{i=1}^{g} m_i$ where the m_i's are the number of observations in group *i* at this stage, the forward search moves to $S^{(m+1)}$ in the following way: after the ANOVA model is fitted to the $S^{(m)}$ subset, *all* the *n* observations are ordered inside each group according to their squared residuals $\tilde{e}_{ij}^2=\left(y_{ij}-\hat{y}_{ij,S^{(m)}}\right)^2$.

For each group *i* we choose the first m_i ordered observations and we add to the *m* observations so chosen the one with the smallest squared residual among the remaining. The ANOVA model is now fitted to $S^{(m+1)}$ and the procedure ends when all the *n* observations are entered the model.

4 The forward search is not sensitive to the method used to select an initial subset, provided unmasked outliers are not included at the start (Atkinson and Riani, 2000, pag 32).

In moving from $S^{(m)}$ to $S^{(m+1)}$, most of the time just one unit joins the previous subset. It may also happen that two or more new units enter $S^{(m+1)}$ as one or more leave, however such an event is quite unusual, occurring only when the search includes one unit that belongs to a cluster of outliers[5].

With the proportional procedure, the only difference is that, given the $S^{(m)}$ subset of dimension *m*, the forward search moves to the next step adding *l* units, where $l=\sum_{i=1}^{g} l_i$ under the condition $l_i/l \approx (n_i - l_{i-1})/n$. Also in this case it may happen (more often than with the previous strategy) that more than *l* new units join the subset as one or more leave.

4.3. *Step 3*: monitoring the search

Both the proportional and non proportional strategies, at each stage of the search, offer the possibility of collecting information on parameter estimates, residual plots, analysis of some statistics of interest, to guide the researcher in the outliers detection and in the analysis of their effect.

If the monitoring of the two strategies shows a different behaviour of the examined statistics, this will be evidence of the presence of different variance in the groups. In fact, with the non proportional strategy, the observations belonging to the groups with the minimum variance will enter before the others. Hence, presumably the outliers will enter the model last, no matter which group they belong to. In the proportional strategy, instead, outliers are forced to enter together with "good" observations in order to maintain the proportionality.

4.4. Non Proportional *vs* Proportional *Forward Search*

In order to illustrate the application and the advantages of the forward search approach we will show the methodology using the sampled dataset described in Figure 1. That dataset is composed by 3 balanced groups generated by a Standard Normal distribution, with the second group heavily contaminated by a 20% of the observations coming from a Uniform distribution *U*(10, 11). As said before, with the classical approach the *F* statistic "erroneously" falls in the rejection area.

5 At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once.

We will illustrate, with the help of graphs, the non proportional approach first, describing the proportional one underlining only the relevant differences and its most important aspects.

Non proportional Forward Search

Figure 2a shows how the observations join the subset $S^{(m)}$ during the search: the subset size is increased by one at each step (x axis). The dotted line lying above the other two refers to group 3 that, by chance, is more homogeneous than the other two groups (see the boxplot in Figure 1): this is why the search algorithm generally choose its units before the others.

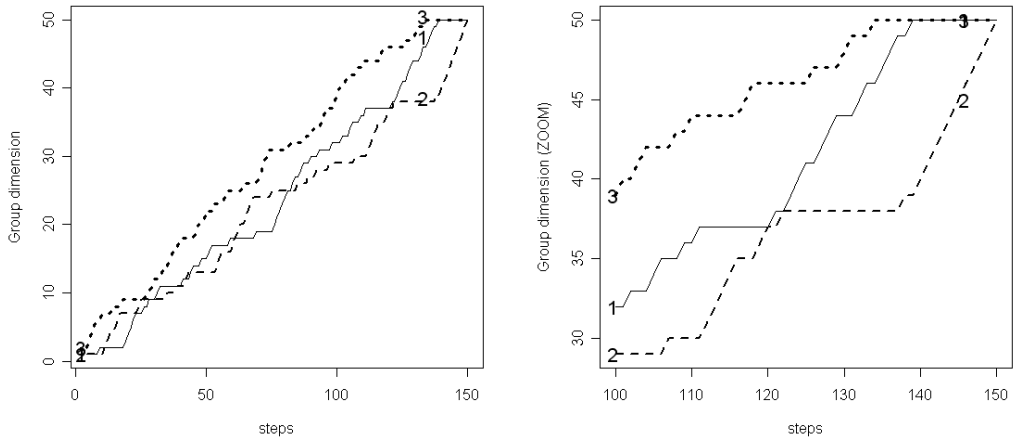

Figure 2. *Plots of the groups dimensions: during the search (a) and zoom of the last 50 steps (b).*

The last 10 observations that join the subset (from step 141 to step 150) belong to group 2 (Figure 2b), giving a first indication of the possible presence of outliers. This behaviour is common to all the samples coming from the sampling design we adopted.

Figure 3 shows how the estimated coefficients change during the forward search. As expected, the estimates vary a lot during the first part of the search because with such a heavy contamination, the medians of the three groups are obviously different and, furthermore, at this stage of the search the effect of the

small dimension of the subsets brings about a high variability of the estimates. As the search goes on, the estimated coefficients stabilize and tend to zero as the number of steps near 140. After that, the estimate of μ_2 increases sharply giving an indication of the presence of outliers in the second group.

Figure 4 shows the *n* residuals computed at each step of the forward search. Throughout the search all the residuals are very small, except for the last 10 entered observations, which are outliers in any fitted subset. Even when they are included in the last steps of the search, their residuals decrease only slightly.

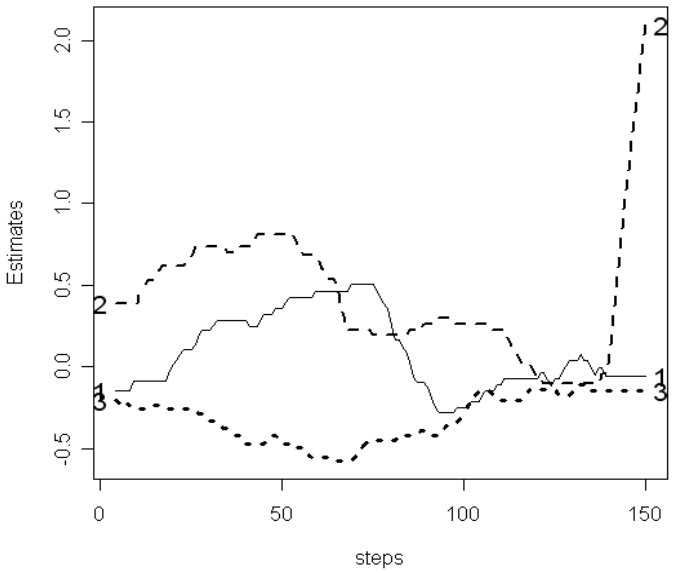

Figure 3. *Forward plots of the estimated coefficients.*

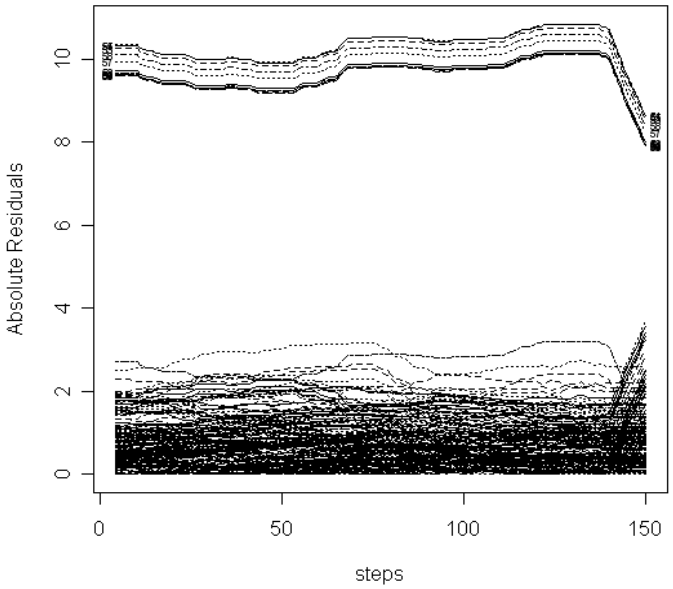

Figure 4. *Forward plots of the absolute residuals.*

Another plot useful to detect the presence of outliers is the one shown in Figure 5a that, at every step of the procedure, reports the changes in the estimate of σ:

$$\hat{\sigma}_m = \sqrt{\frac{\sum_{i=1}^{m} e^2_{i,S^{(m)}}}{m-g}}$$

where $e^2_{i,S^{(m)}}$ are the m squared residuals of the subset $S^{(m)}$. Initially $\hat{\sigma}_m$ is close to zero, underestimating the "real" parameter $(\sigma = 1)$ because of the selection rule. Obviously, the inclusion of each further unit causes an increase in the value of the estimate, that will tend to 1 as the number of observations tends to 140. After that, the estimate increases sharply giving a plot that is virtually in the form of two line segments, one for each group of observations (non outliers and outliers). The monotone form of this plot indicates that the observations have been correctly sorted by the forward search. This plot, together with the ones of the estimates and of residuals allows the identification of heteroschedasticity even under H_0. In case of heteroschedasticity with no contamination, the estimates will

assume values around the true mean without big jumps during the search, while the residual standard error will increase (more or less sharply depending on the strength of the heteroschedasticity) in the final steps.

It is also possible to monitor (see Figure 5b) the *t* statistics relating to the coefficients:

$$t_{i,S^{(m)}} = \frac{\hat{\mu}_{i,S^{(m)}}}{s.e.(\hat{\mu}_{i,S^{(m)}})}$$

where $\hat{\mu}_{i,S^{(m)}}$ is the estimate of the mean for the group *i* at step *m*.

Initially, when the standard error is close to zero, the statistics are very large and off the scale of the plot. As the process progresses, the variance of the observations becomes larger and the absolute value of the statistics $t_{i,S^{(m)}}$ keep decreasing until the end of the search if there is no contamination. If one or more groups are contaminated, instead, toward the end of the procedure there should be a sharp increase in one or more of the absolute values of the *t* statistics.

Figure 5b reports also the 95% acceptance region (grey lines) of the three *t* tests only for step $m \geq 30$, when the *t* can be suitably approximated by the Standard Normal distribution.

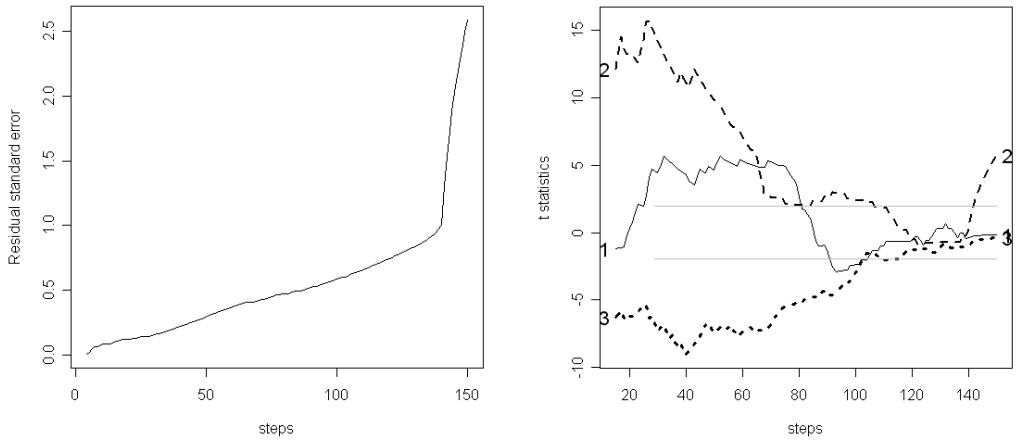

Figure 5. *The increasing value of the residual standard error (RSE) and of the t statistics.*

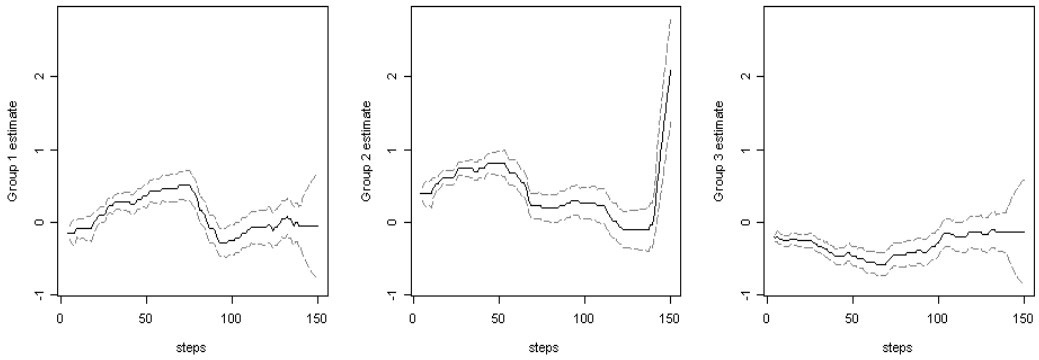

Figure 6. *Forward plots of the estimated coefficients with their significance band.*

Figure 6 shows the three plots of the coefficient estimates relating to their confidence interval at significance level of 5% (grey broken lines). At the end of the search, as $\hat{\sigma}^2$ increases because of the introduction of observations with large variance, the bands become larger.

The computation of $R^2 = 1 - \frac{DW}{DT}$ gives us further indication of the presence of outliers (see Figure 7). In the graph, broken lines represent at each step the 5% and 95% quantiles of the empirical $R^2_{S^{(m)}}$ distribution, obtained from a Montecarlo simulation of 10000 samples free of contamination. Increasing *m*, $R^2_{S^{(m)}}$ decreases to zero showing group means equality. Only in the last few steps, the value of the statistic strongly increases in accordance with the change in the parameter estimates. Again, this shows that the last observations must be outliers.

Figure 8 shows how the *F* statistic moves during the forward search. The grey lines represents the 95% and 99% quantiles of the Fisher *F* distribution with 2 and $m-3$ degrees of freedom at each step, while the black one refers to the values of $F_{S^{(m)}}$ obtained during the procedure. Initially, these values are very large, all in the rejection region and often off the scale of the plot. After a while, the statistic decreases and, for $112 \leq m \leq 142$, falls in the acceptance region.

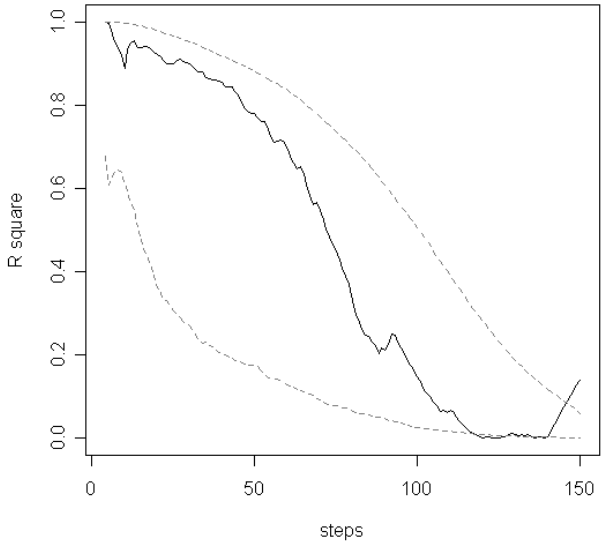

Figure 7: *Forward plot of the R^2 . Dotted lines are the 5% and 95% quantile of the R^2 empirical distribution obtained from a simulation study without contamination.*

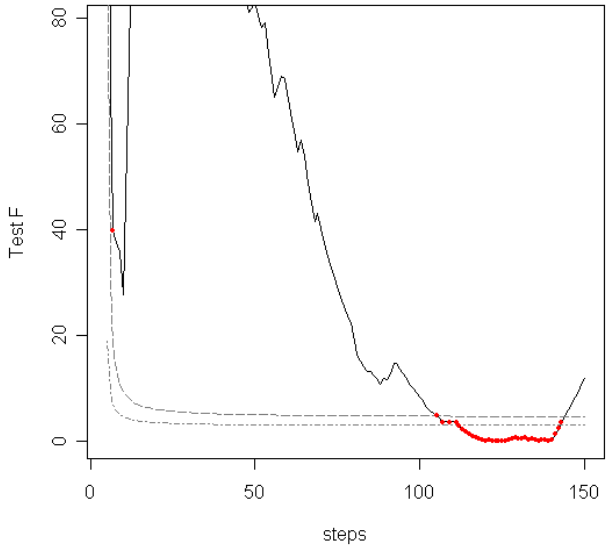

Figure 8. *Forward plot of the test F. Broken grey lines identify the acceptance region at 5% and 1% levels of significance.*

As the forward search proceeds, the F value suddenly increases above the critical value and keeps increasing till the end of the search. At the last step, the value of the F is 12.001, as already showed in paragraph 3.

Proportional Forward Search

With the proportional forward approach, observations join the subset $S^{(m)}$ in a proportional way: forward plot of group dimensions are approximately straight lines.

Figure 9a describes the estimated coefficients at each step of the search; in this case the graph is very similar to the graph shown in Figure 4: the estimates are different during the first part of the search, then they tend to zero becoming quite close to each other; in the last few steps, only the estimate of μ_2 increases widely giving an indication of the presence of outliers in the second group.

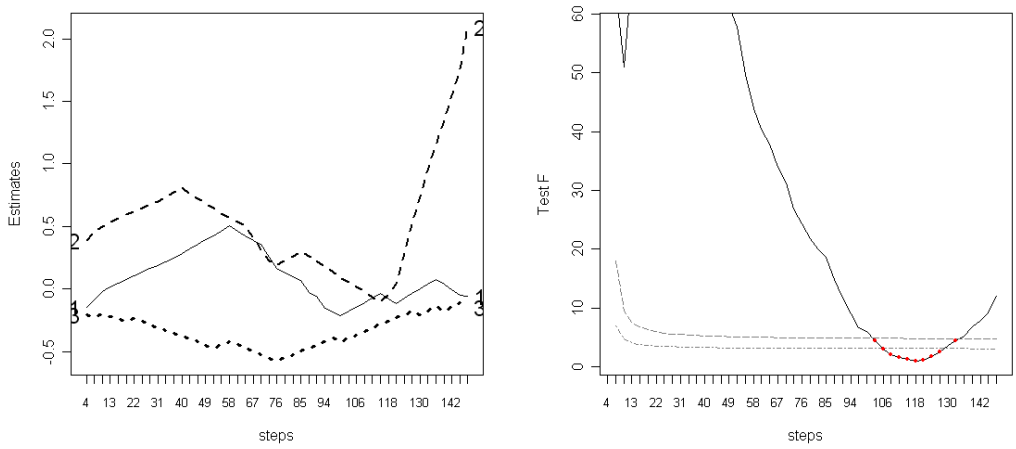

Figure 9. *Forward plots of the estimated coefficients and of the test F.*

Obviously, with the proportional approach, $\hat{\mu}_2$ starts to increase before $m=140$; this is due to the fact that the proportional forward algorithm enters l_i units for group *i* at each step (a part from the difference due to the rounding to the closest integer), and the outliers join the subset before the non proportional approach.

Figure 9b shows how the *F* statistic moves during the proportional forward search. Obviously, the only difference with respect to the graph of Figure 8 (referring to the non proportional search) is that the *F* statistic enters and exits the acceptance region bands (grey line) at the step $m = 103$ and $m = 130$ while the non proportional approach gave $m = 112$ and $m = 142$.

5. Forward *F* test and its evaluation.

The previous paragraph makes clear the utility of the forward search in the ANOVA in detecting the possible presence of outliers and their effect on parameter estimates and on aspects of inference about models. As showed, the search is completely based on graphical analysis and is strictly connected with the contest of the observed phenomenon.

The forward test F can be defined as a collection $F_{FS}=\{F_{(k)},\ldots,F_{(n)}\}$ of the classical F test in each step of the search (see Figure 8 and 9); to obtain a robust forward F test it is possible to individuate a cut-off point of the progress procedure dividing the group of observations that differ to the bulk of the data from the others. Naturally, the search of the cut-off point can not be "automatic" but depends on the complete analysis and knowledge of the phenomenon. In this context, the robustness of the method does not derive from the choice of a particular estimator with a high breakdown point but from the progressive inclusion of units into a subset which in the first steps is outlier free.

As mentioned before, the cut-off point is chosen by a thorough examination of the graphs presented in the previous section and on other *ad-hoc* considerations. It is clearly impossible to carry out a simulation to evaluate the goodness of the test without an "automatic" cut-off point. To give an idea on how the robust test behaves we refer to the simulated example illustrated in the paragraph 3. Datasets are composed by three balanced groups composed by n_i $(n=20,40,\ldots,200\ ;\ i=1,2,3)$ observations coming from a Standard Normal distribution. Only the distribution of the second group is $(1-\epsilon)N(0,1)+\epsilon N(2,1)$ where $\epsilon=0.05,\ldots,0.10$. We decided to stop the search procedure (cut-off point) at $\epsilon\cdot n_i$ steps before the end of the forward search and to use the non proportional approach because it guarantees that the outliers enter the subset at the end of the search. We then compared the results with the ones obtained with the classical approach.

Table 2 shows frequencies over 10000 simulations in which robust forward F test falls in the rejection area at the nominal significance level of $\alpha=0.05$. For example, for the pair $(n_i=100,\ \epsilon=0.08)$ our forward F test produces an evidence versus the alternative hypothesis in 1160 samples while the for the classical test the frequency is of 1629 over 10000 replicates. With the proposed method, the probability of accept H_1 when H_0 is true is always lower than the same probability showed in Table 1. However, the frequencies relating to the higher values of ϵ are quite high because of the automatic procedure used to set the cut-off point; in particular, we did not analyse every sample with a graphical approach as the forward search suggest.

	$\epsilon =$	5%	6%	7%	8%	9%	10%
$n_i =$	**20**	0.0453	0.0453	0.0453	0.0491	0.0491	0.0491
	40	0.0434	0.0434	0.0591	0.0591	0.0629	0.0629
	60	0.0573	0.0532	0.0532	0.0827	0.0827	0.0818
	80	0.0478	0.0741	0.067	0.067	0.1089	0.0955
	100	0.0686	0.0598	0.0972	0.0823	0.1354	0.1175
	120	0.0519	0.0875	0.0708	0.0976	0.1598	0.1341
	140	0.0828	0.0669	0.0864	0.146	0.1876	0.1442
	160	0.058	0.0729	0.1277	0.1646	0.124	0.1569
	180	0.0935	0.1158	0.1522	0.1119	0.1405	0.1703
	200	0.0669	0.0828	0.1003	0.1273	0.1512	0.1850

Table 2. *Approximation of the true type I error probability of the Forward robust F test (non proportional approach).*

6. An evaluation of the Italian University reform

As an application of the proposed approach to real data, we use a set of information referring to the performance of the Italian University System. The data come from annual surveys conducted by the Italian National University Evaluation Committee (NUEC) during the past five years (2001 – 2005) and refer to the activities of all the public universities during the academic years 1999/00 - 2003/04. The collected information are used to compute a set of 29 indicators grouped in 4 classes (*Outcome*, *Resources*, *Process* and *Contextual* indicators).

For our purposes, we decided to use the first year *retention rate* indicator *RT*, (a *Process* indicator defined as: $RT = 1 - WR$ where *WR* is the withdrawal rate) to evaluate the impact on the Italian university system of the reform on degree programs, that was enacted in the academic year 2001/02 (Bini *et al.*, 2003) . Our procedure is applied to a model in which *RT* is the dependent variable and, as an example, we limited our analysis to the Italian degree programs in Mathematical Science. With the ANOVA model we want to find out the effect of the reform on the retention rate over different years. Since in our data we found many anomalous observations, our procedure is indicated for the estimation of the model.

The data of the year 2003 (the academic year in which the reform took place) are not used in our analysis since they are affected by too many collection errors. Hence our dataset is composed of four groups identified by the years in which the NUEC surveys were conducted (the two academic years before and after 2003), composed respectively of 276, 283, 342, 351 observations. On this dataset we conducted a classical ANOVA test whose results are shown in Table 3. The *F* value falls in the accepting region; from this analysis we could then say that the reform had no effect on the first year retention rate.

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	*F* value	Pr(>F)
f	3	0.18034	0.06011	2.50055	0.058018
Residuals	1248	30.00779	0.02404		

Table 3. *Analysis of Variance Table on the RT index of the Italian degree programs in Mathematical Science over the NUEC surveys years 2001, 2002, 2004 e 2005.*

An analysis of the boxplot for the four groups (see Figure 10) shows the presence of observations that seem to differ from the bulk of the data in each group and that could be considered outliers.

It was decided then to carry out our robust procedure. As is clearly shown in Figure 11b, these observations enter the model at the end of the search. Since outliers are especially numerous in the groups 2004 and 2005, the slopes referring to these years are steeper than the others.

The estimates of the coefficients at each step are shown in Figure 12. The values referring to 2004 and 2005 are always higher than the others, showing a clear evidence of a positive effect of the reform on the first year retention rate *RT*. The outliers effect is evident since the estimates of these years converge to the others only in the last steps of the search.

The presence outliers is very clear if we analyse the plots of the absolute residuals and of the residual standard error that increases in an exponential way during the last steps of the search (Figure 13a and 13b).

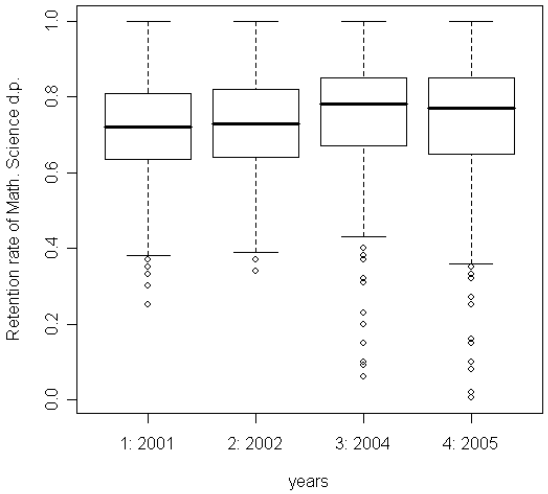

Figure 10. *Boxplot of the first year RT index computed for all the Italians Mathematical Science degree programs.*

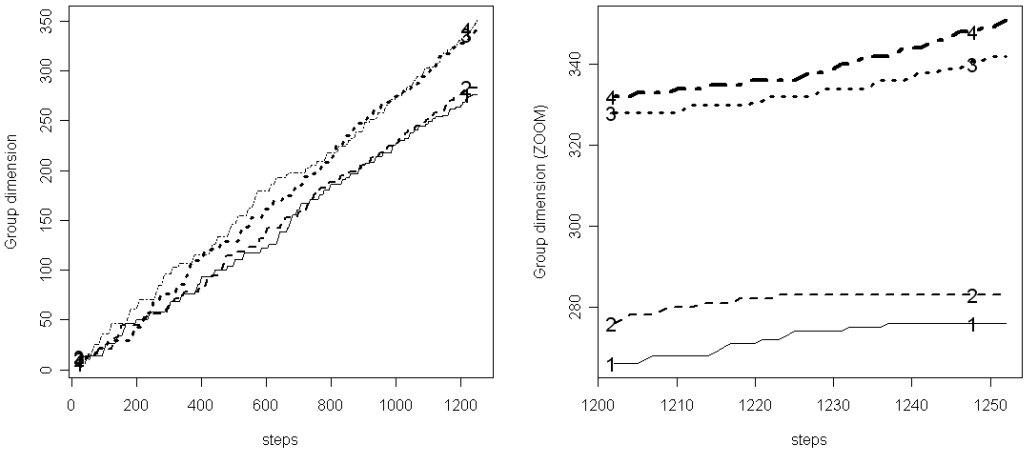

Figure 11. *Plots of the groups dimensions: during the search (a) and zoom of the last 50 steps (b).*

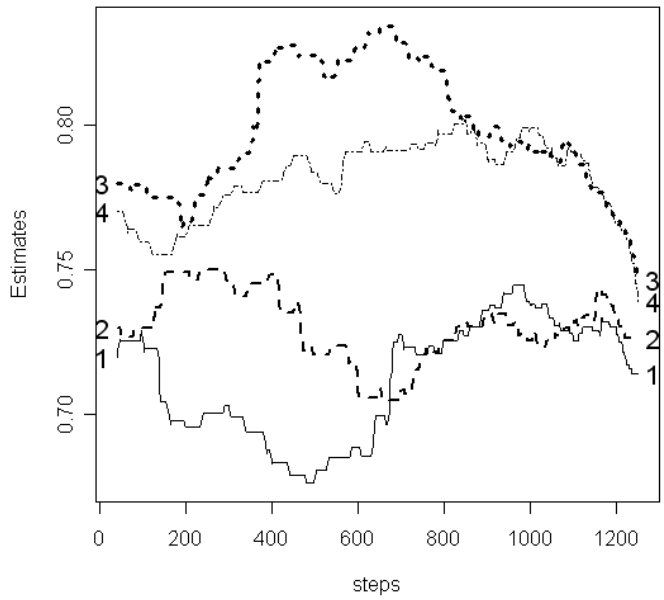

Figure 12. *Forward plots of the estimated coefficients for the first year retention rate of the Italian degree programs in Mathematical Science.*

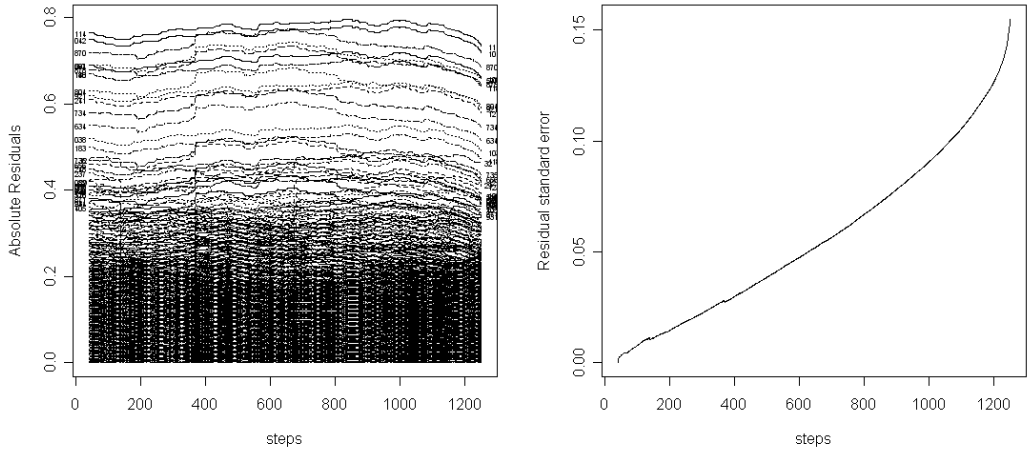

Figure 13. *NUEC data: forward plots of the absolute residuals and of the residual standard error.*

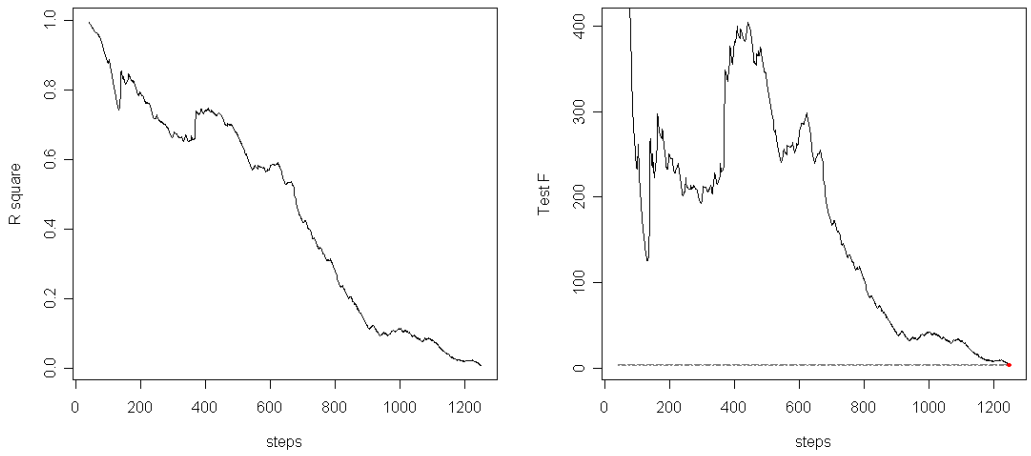

Figure 14. *NUEC data: forward plots of the* R^2 *and of the F statistics. In the last steps the* R^2 *tends toward zero and the F enters in the acceptance region.*

Finally, let us analyse the graphs in Figure 14a and 14b that describe the behaviour of the R^2 and of the F statistics. Again, the presence of outliers is evident by the steeper decrease of the two statistics during the last steps of the search. In the last two steps the F value fall in the acceptance region of H_0, in accordance to the result of Table 3.

As mentioned in the previous paragraphs, our procedure is based on a graphical approach. Controlling the outliers effect bring us to conclude for a positive effect of the reform, contrary to the classical approach result.

Concluding remarks

One of the most important topic in statistical inference is the presence of outliers in the data. Our work concerns the effect of outliers in the ANalysis Of VAriance: this methodology is very powerful under classical assumptions, but it is strongly affected by the presence of outliers. We implemented the *Forward Search* method in the ANOVA framework, in order to individuate the observations that differ from the bulk of the data and to analyse their effect on the estimation of parameters and on inferences on the model.

The methodology proposed takes into consideration the presence of groups in the data structure of the model. At every step of the forward search we compute parameters estimates, residuals, classical *F* values and other considerable statistics. We implemented two approaches to carry on the analysis: proportional and non proportional; the difference is in the number of units that join the model during the search and points out some characteristics of the data structure. Finally, we proposed a procedure to obtain a robust forward *F* test individuating, with a graphical approach, a cut-off point of the classical *F* test values in each step of the search that divides the group of outliers from the other observations. We derived from a Montecarlo simulation study that with the proposed method the probability of the the *type I* error is lower than with the classical ANOVA.

In order to illustrate the application and the advantages of the forward search approach we used some artificial examples. Furthermore, we showed an application of the proposed approach to real data, using a set of information referring to the performance of the Italian university system. The data come from annual surveys conducted by the Italian National University Evaluation Committee. We used the first year *retention rate* indicator to evaluate the impact on the Italian university system of the reform on degree programs, enacted in the academic year 2001/02. Our procedure is applied to a model in which *RT* is the dependent variable; with the ANalysis Of VAriance we want to find out the effect of the reform on the retention rate over different years. Since in our data we found many anomalous observations, our procedure is indicated for the estimation of the model. Despite to the presence of a big number of outliers in the last two years of the NUEC survey, the application shows the positive effect of the Italian University reform in reducing the withdrawal rate.

References

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

Atkinson A. C., Riani M., Cerioli A. (2004). *Exploring Multivariate Data with the Forward Search.* Springer, New York.

Barnett V. (1988). “Outlier and order statistics”. *Commun. Statist. Theor. Meth.* **17**, 2109-2118.

Barnett V., Lewis T. (1993). *Outliers in Statistical Data (3rd edition)*. Wiley, New York.

Bini M., Bertaccini B., Polverini F. (2003) The use of outliers for the evaluation of public policy activities: the first year college drop out rate in Florence. *Proceedings of the Joint Statistical Meetings of the American Statistical Association.* August 2003, San Francisco.

Hampel *et al*. (1986). *Robust statistics: the approach based on influence functions*. Wiley, New York.

Olive D.J., (2005). *Applied Robust Statistics*.
On-line book available at http://www.math.siu.edu/olive/ol-bookp.htm.

Staudte R.G., Sheather S.J. (1990). *Robust estimation and testing*. Wiley, New York.

Tukey J. W. (1960). “A survey of sampling from contaminated distribution”. In Olkin, I. (1960). *Contributions to Probability and Statistics*. University Press, Stanford, California.