Chapter 9
Robust diagnostics in university performance studies

Matilde Bini, Bruno Bertaccini and Silvia Bacci

> *In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions, and the introduction of which into the investigations can only serve to perplex and mislead the inquirer.*
>
> (Peirce, 1852)

9.1 Introduction

The presence of anomalous observations (*outliers*) in a set of data is one of the greatest problems in methodological statistics, one that scientists were already aware of many years ago, as can be seen in the comments made by the American astronomer Peirce[1] over 150 years ago.

An anomalous value can be defined generally as an *observation that appears to be non-compatible with the probabilistic model that has generated the rest of the data* (Barnett and Lewis, 1993). However, this statement bears a certain degree of subjectivity the moment the judgement of compatibility has to be expressed. This concept can be explained if we consider certain divorce cases that have been filed on the grounds of abnormalities discovered in the duration of pregnancies (Barnett, 1976). In 1949 a certain Mr. Hadlum appealed against the rejection of his previous

Matilde Bini
Department of Statistics “G. Parenti”, University of Florence, Viale Morgagni 59, 50134, Florence, Italy, e-mail: bini@ds.unifi.it

Bruno Bertaccini
Department of Statistics “G. Parenti”, University of Florence, Viale Morgagni 59, 50134, Florence, Italy, e-mail: brunob@ds.unifi.it

Silvia Bacci
Department of Statistics “G. Parenti”, University of Florence, Viale Morgagni 59, 50134, Florence, Italy, e-mail: s.bacci@ds.unifi.it

[1] Peirce was the first to test for the objective identification of anomalous observations.

M. Bini et al. (eds.), *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, Contributions to Statistics,
DOI 10.1007/978-3-7908-2385-1_9, © Springer-Verlag Berlin Heidelberg 2009

case for divorce, which was based on his wife's suspected adultery because she had given birth to a child 349 days after he, the husband, had left to go to the war. Since the average gestation in humans is 280 days, 349 days seemed surprisingly long to Mr. Hadlum, causing him to judge the span of time an anomalous value, that is, an observation deriving from another population (in this case, originating after the moment declared by the wife). Mr. Hadlum lost the case; the Court of Appeal maintained that the wife's period of gestation, while highly improbable ("extreme" value), was not scientifically impossible, in contrast to Mr. Hadlum who considered the value "contaminant", that is, deriving from a different distribution and therefore clear evidence of the wife's adultery. It is very likely that Mr. Hadlum would not have been suspicious if, for example, his wife had given birth 290 days after his departure for the war: nevertheless, even this length of time for the gestation could have been considered a "contaminant" value in the afore-mentioned sense, since the wife could have become pregnant, for instance, 20 days after the husband's departure plus a normal gestation of 270 days.

The keywords used here are *outlier, extreme value and contaminant value*. Therefore, these concepts must be made clearer, and an attempt must be made to formalise their definitions; for this purpose, we must consider an ordinate sample of n univariate observations

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

all deriving from a certain F distribution, with the exception of two from a G distribution (see Fig. 9.1).

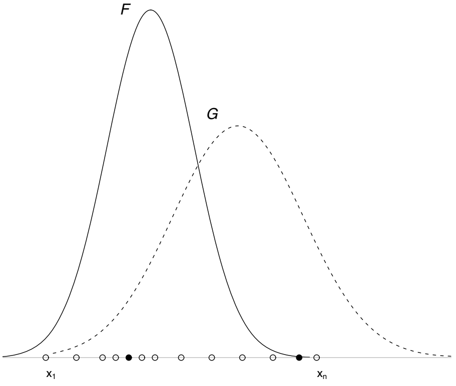


Fig. 9.1 Outliers, extremes values and contaminants.

The $x_{(1)}$ and $x_{(n)}$ observations are the extreme values of the sample; in the case being examined, however, only $x_{(n)}$ can be considered anomalous because of its position in relationship to the F model generator hypothesized. Hence, the extreme values are not necessarily *outliers*, whereas any individual *outlier* is always an extreme (or relatively extreme) value of the sample. The observations deriving from the G distribution are indicated with a black dot in the figure; though both can be defined contaminant values, only the second one, coinciding with $x_{(n-1)}$, can be considered anomalous regards F, differing from $x_{(n)}$ (even more anomalous than $x_{(n-1)}$), which nevertheless is not contaminant. Hence, the contaminant values can be or cannot be identified as *outliers*, and these can be more or less contaminant values (that is, deriving from distributions different to the one hypothesized). Unfortunately, there is no way we can know if an observation is contaminant (this is why Mr. Hadlum would probably not have been suspicious of a pregnancy lasting 290 days); all we can do is try to understand *if the outliers are possible manifestations of some form of contamination* (Barnett, 1988).

The importance of this problem led Box and Andersen (1955) to coin the term "robust" when referring to estimation methods that continue to have desirable properties, in spite of the fact that part of the data might result presumably contaminated to a certain degree.

In this respect, Tukey (1960) defines the mixture $(1-\varepsilon)F+\varepsilon G$ a contaminated distribution, where the F distribution is contaminated by the G distribution with ε probability (known as *contamination quota*). In his famous work, for evaluating the effects of a casual-type contamination on the efficiency properties of traditional estimation procedures, Tukey presumed the extraction of a sample of n observations from the contaminated distribution

$$(1-\varepsilon)N\left(\mu,\sigma^2\right)+\varepsilon N\left(\mu,9\sigma^2\right).$$

Figure 9.2 illustrates the effect of the contamination Tukey proposed for certain ε values between 0 and 0.5: the contamination weighs down the tails of the original distribution, and this weight becomes heavier as the ε value increases.

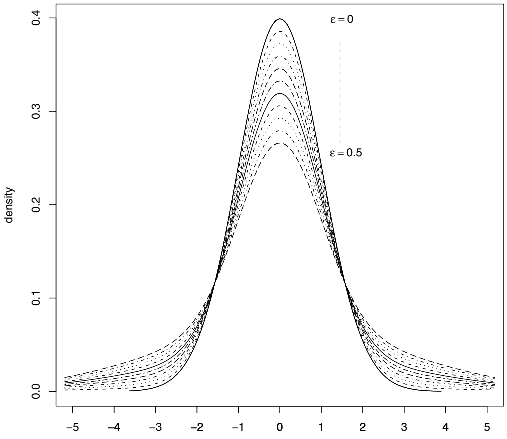


Fig. 9.2 Tukeys' mixture.

Tukey demonstrated how small, natural contaminations (between 1 and 10%) in the theoretic model could make the traditional asymptotic theory on optimality absolutely insignificant.[2]

There are various sources of contamination that can produce anomalous values; sometimes they are concomitant and they are certainly never known beforehand. The anomalous aspect might reflect the natural variability of the phenomenon being investigated, generated by erroneous measurement or, more often, they might derive from mistakes in the "implementation of the design" caused by distraction or due to ignorance of the person responsible for recording the information.[3] But detection of these sources appears as a problem of secondary importance compared to the adoption of tools that allow efficacious identification of the *outliers*.

After a description of the characteristics of robust methods in Sect. 9.2, an introduction of the Forward Search algorithm and its implementation in the Generalized Linear Models applied to the university effectiveness evaluation are reported in Sect. 9.3. Section 9.4 presents the Forward Search for the fixed effects ANOVA models and its application to the evaluation of the Italian University reform. Finally, Sect. 9.5 is devoted to some concluding remarks.

9.2 Robust methods vs diagnostic analysis

The aim of any diagnostic analysis is to define a more or less general outline of the collective phenomenon under investigation, in order to identify the peculiar characteristics, which, at a further stage, will require advanced statistical techniques and tools. This process of comprehension often originates in the implementation of a *probabilistic model*,[4] capable of synthesizing the state of knowledge of the phenomenon in question. Therefore, the model is a description of the process that generates the data and its implementation becomes the main criterion for recognition of the anomalous observations; hence, a non-typical value is considered such if it is able to produce a "surprise" effect regarding the particular probabilistic model presumed to have generated the data. For instance, in a sample of 7 observations made up of:

$$0.47, 6.18, 0.09, -0.60, -1.09, -1.19, 1.86$$

the second value is surprising if connected with a theory that presumes a probabilistic model of the N family $(0,1)$ to be the generator; however, this supposition

[2] In his famous work, Tukey proposed a comparison of relative efficiency between the $d_n = \frac{1}{n}\sum|x_i - \bar{x}|$ and $s_n = \left[\frac{1}{n}\sum(x_i - \bar{x})^2\right]^{1/2}$ estimators of the σ variability parameter, revealing how only two out 1,000 observations are able to annihilate the efficiency of the s_n estimator; in particular, Tukey demonstrated how d_n results to be preferable for all the ε values within the interval [0.002, 0.52].

[3] The mistakes in measurement and in "implementation of the design" (which lead to the inclusion of non-representative units) are defined by Anscombe (1960) as false observations.

[4] The description of any type of reality is often a very complicated operation, because of the interrelationships that are nearly always present.

would appear completely misleading since the data have been really generated by a Cauchy distribution with a parameter of 1 on the scale.[5]

From a typically parametric point of view, the initial theories indicate a probabilistic model in this manner, hoping it to be only a fair approximation of reality, without ever being able to presume that it is absolutely correct. Hence, all statistic procedures should have the following, desirable properties (Huber, 1981):

1. they should demonstrate a reasonable (almost optimal) level of efficiency regards the model presumed;
2. they must be robust, meaning that slight deviations from the theoretic model should bring about similarly slight penalization in performance (for instance, the asymptotic variance of an estimator ought to be near its nominal value as calculated in relationship to the theoretic model);
3. any appreciable deviation from the theoretic model should not cause a "catastrophe".

At this point, one might wonder if the robust procedures are really necessary or if, on the other hand, it would be sufficient to resort to the traditional procedures after adopting some technique that can discard the anomalous observations.

Unfortunately, this is not the case. First of all, the techniques that discard outliers are not free of errors; in Tukey's example, the removal of *outliers* from the dataset generated by mixed distribution would continue to produce a sample of observations that are not normal – because of wrong exclusions and wrong preservation of data – leading to a *framework* that would be just as inappropriate as the initial one, which advises against applying the traditional theory to normality.

Moreover, the difficulties encountered in the detection of anomalous values increase as the number of variables composing the structure of the available data rise. In fact, a univariate analysis of the context, while being an important part of the statistical procedure, is often of limited interest since many modern investigational techniques (confirmed by appropriate graphical analyses) are able to distinguish atypical situations.

Much more interesting is the multivariate analysis, in which the spatial composition of the observations makes the placing of anomalies less intuitive and, consequently, the formal methods for detecting them much more complex.

Lastly, the best procedure for removing *outliers* cannot match the performance shown by the best robust procedure Barnett and Lewis (1993). Indeed, this latter is definitely superior because it is a gradual (and not immediate) transaction between the total acceptance and the total deletion of an anomalous observation manifesting a contaminant distribution.

In the case of linear regression models, the presence of anomalous values can be easily depicted by simply *plotting* the data or the residuals; however, even in the multiple regression model, when the number of explicative variables increases, their detection by means of graphical tools may not be so immediate, especially in the presence of groups of *outliers* that mask each other.

[5] Observations beyond the body of the data can be caused by casual extractions from the Pareto and Cauchy distributions.

To overcome this problem, some estimation methods called *robust* or *resistant* have been proposed; these terms are used in the literature to illustrate their capacity to produce estimates that are not easily influenced by contaminant data. These methods all identify as *outliers* those units that show the highest residuals. Among the various *robust* approaches proposed, special mention must be made of the Least Median of Squares – LMS – (Rousseeuw, 1984) because it is intuitive and easy to use. However, the robust estimators (*LMS, MAD, trimmed mean*, etc.) have the disadvantage of under weighing or neglecting some of the observations; furthermore, they can fail completely if the observations do not derive from one population alone, but from various distinct populations.

Another approach to the problem is through the so-called *diagnostic analyses,* which foresee statistical calculation capable of detecting the anomalous values and the most influent among them. These can be examined and then either deleted or corrected, in order to allow the model to re-adapt by means of the traditional techniques. Worthy of note among the diagnostic techniques is that known as the *single deletion diagnostic*, which, at every step of the analysis, foresees the elimination of one observation at a time from the *n* available, and calculation of the new estimates and new parameters on the remaining observations. With two *outliers*, pairs of observations can be deleted and the process can be extended to several units at the same time. However, the traditional diagnostic methods suffer from serious inconveniences; the masking effect that takes place in the presence of groups of *outliers* makes the individual influence of each single one very limited and therefore unidentifiable; this aspect requires the diagnostic process to be extended to several observations simultaneously. Nevertheless, one realises immediately that the combinatory explosion of the number of observations to be taken into account can create considerable problems from a computational point of view and in interpretation of the results. An alternative to these limitations can be to repeat the single deletion processes; but in this case, the set of observations used in the adaption process decreases as the analysis proceeds. Atkinson and Riani (2000) demonstrated that this procedure, which is commonly known as *backward deletion*, might sometimes fail.

Note that the *robust* approaches, in spite of the fact that they focus on the same diagnostic target, proceed in a completely opposite manner, adapting the model first of all by using techniques that take into account the characteristics of the dataset, then examining the units that diverge most from the predicted values. However, the two approaches often lead to the same results. Some authors, while agreeing on the need to resort to robust criteria for the analysis, do not approve of the deletion of cases that have been really observed (though many robust methods do not consider the *outliers* at all); on the contrary, others, in spite of agreeing on the necessity to remove the anomalies, maintain that to resort to a robust method rather than another type is arbitrary (even though the preventive deletion of the observations and the subsequent adaption by means of ordinary least squares is itself a robust method).

In effect, this debate does not at all solve the problem of the *outliers*. What is undoubtedly important is to judge each single technique on the basis of the number of *outliers* it manages to identify – or tolerate – before they can influence the inferential process somehow. This property is formalised by introducing the *breakdown point* (Donoho and Huber, 1983; Rousseeuw and Leroy, 1987) which is defined the smallest fraction of contamination that can make a certain estimator assume values far away from the estimates that would have been obtained if the contamination was absent.[6]

9.2.1 The Forward Search algorithm

The *Forward Search* (Atkinson and Riani, 2000; Atkinson et al., 2004) is a procedure that is capable of combining the efficiency of traditional inferential methods with the capacity to identify anomalous observations within a sample of data and then assess the effects achieved. Its main feature is that it proceeds in a manner exactly opposite the *backward* one, that is typical of the traditional diagnostic methods that assess the anomaly or the influence of an observation on the statistic model only after this has been adapted to the entire sample of data. On the contrary, the *Forward Search,* given the n observations of the sample, starts by searching among the data available for a minimal dataset presumed – on the basis of the model – to be free from *outliers*. This starting subset is detected by means of different approaches according to the analysis context: in the case of linear regression models, the adaption of a high number of small subsets is evaluated, employing robust statistical methods to define which of these procedures produces the best adaption; in the case of multivariate statistical analysis, *boxplot* bivariate matrix and *spline* functions adapted to the actual placement of the observations in space are used.

The evolution of the procedure is therefore ensured by evaluation of the adaption to increasingly larger observations obtained by the sequential inclusion of the remaining observations in relationship to their proximity to the theoretic model. The process obviously stops when all the units observed participate in the inferential process. The arrangement of the data performed at every step excludes the problems of masking encountered by the traditional diagnostic methods, the targets of which are reached by monitoring the various statistics (i.e. the goodness of fit test, significance of parameters) during the evolution of the algorithm.

The result of this procedure is the arrangement of the observations with respect to the degree of their proximity to the presumed model[7]; in the case of linear regression, this arrangement is achieved by starting from a robust adaption and reaching one of ordinary least squares.[8] Monitoring the various statistics usually employed in the traditional inferential approaches permits gathering a set of data capable not only of detecting the *outliers* but also – and this is even more important – of understanding the influence that each of them has on the inference of the model.

The next sections propose the robust diagnostic analysis made using the forward search algorithm as a useful tool to detect anomalous situations in university performance evaluation.

[6] For example, for ordinary least squares even one observation alone might be sufficient for this to take place. In this case, the OLS estimator would have a breakdown point of 0%.

[7] In regression models, “proximity” is expressed by the residuals; in multivariate analysis, by a measurement of distance (Mahalanobis, Manhattan, etc.).

[8] If the model agrees with the data, the robust adaption and the least squared one will produce similar results, both in estimation of the parameters and in the errors. However, the estimates and the residuals of the adapted model change considerable during the search process.

9.3 The Forward Search for Generalized Linear Models

The Forward Search is an approach for detecting the presence of outliers and assessing their influence on the estimates of the model parameters. The method was first applied to regression analysis, but it could as well be applied to almost any model (Atkinson et al., 2004). The procedure starts out by fitting the model to a subset of the observations, say m observations, which is chosen in some robust way. The observations of the entire set are then ordered by their closeness to the estimated model. The model is then refitted using the subset of the $(m+1)$ observations which are closest to the previously estimated model. The observations are ordered again, the model is refitted to a larger subset and the process is continued until all the data have entered. At every step the subset size is increased by one unit (usually one case is added to the previous subset, but sometimes two or more are added as one or more leave the subset), bringing about an ordering of all the observations. At every step, the fitting of the model will also produce estimates of the parameters of the model under study as well as other relevant statistics. Changes of these statistics, as the Forward Search is carried out, are analyzed (graphically or otherwise) for the purpose of assessing the influence of each observation on the estimation of the model and - under the hypothesis that the outliers are the last ones to enter - of identifying a cut-off point that divides the outliers from the “good” data. More formally, it is based on the following steps:

The start is a robust fit to very few observations and then a successive fit is done with larger subsets. The initial subset is identified using the *least median of squares method* (Rousseeuw, 1984) that guarantees that no outliers are included in the initial subset.

Formally, (see details in Atkinson and Riani, 2000, p. 31): let $Z = (X,y)$ a data matrix of dimension $n \times (p+1)$. If n is moderate and $p << n$ the choice of the initial subset can be performed by exhaustive enumeration of all $\binom{n}{p}$ distinct $p$tuple $S^{(p)}_{i_1,\dots,i_p} \equiv \{z_{i_1},\dots,z_{i_p}\}$, where $z^T_{i_j}$ is the i_jth row of Z, for $j = 1,\dots,p$ and $1 \le i_j \ne i_{j*} \le n$.

Specifically, let $\iota^T = [i_1,...,i_p]$ and let $e_{i,S^{(p)}_\iota}$ be the least squares residual for the unit i given the model has been fitted with the observations in $S^{(p)}_\iota$. The initial subset is $S^{(p)}_*$ which satisfies

$$e^2_{[\mathrm{med}],S^{(p)}_*} = \min_\iota \left[e^2_{[\mathrm{med}],S^{(p)}_\iota} \right] \tag{9.1}$$

where $e^2_{[k],S^{(p)}_\iota}$ is the kth ordered squared residual among $e^2_{i,S^{(p)}_\iota}$, with i=1.. . . .n and med=integer part of $(n+p+1)/2$. If $\binom{n}{p}$ is too large, the choice is made using 3,000 $p$tuples sampled from Z matrix.

The subset size is increased by one and the model refitted to the observations with the smallest residuals for the increased subset size.

The initial subset $S_*^{(m)}$ of dimension $m \geq p$ is increased by one and the new subset $S_*^{(m+1)}$ consists of $m+1$ units with the smallest ordered residuals $e^2_{[k],S_*^{(m)}}$. The model is refitted to the new subset and the procedure continues increasing subset sizes until all the data are fitted, i.e. when $S_*^{(m)} = S^{(n)}$.

The result is an ordering of the observations by closeness to the assumed model.

9.3.1 Robust GLMs for the university effectiveness evaluation. The case of the first year college drop out rate

One of the most important indicators of efficacy, as well as of efficiency, which the Ministry takes into account in judging the teaching activity of a university, is the *drop out rate*, which continues to be extremely high in all the Italian universities even after the reform. Recent findings confirm that the drop out rate is still well over 30 percent and this calls for new research to discover the reasons and possible solutions of a problem that has strong social and political implications.

Past research and commonsense tell us that the most important factors affecting the probability of dropping out from college are the characteristics of students at time of enrollment, but also possible changes of their characteristics during their study, as well as the characteristics of teaching activity. These factors are important for every university but their impact is probably different for different institutions. This justifies a research on drop out rate conducted on data from the University of Florence (Italy) (Bini et al., 2003; Bini and Bertaccini, 2007).

The present study has been performed using two sets of data which have been linked. Administrative data, collected by each Italian university at time of enrollment and survey data, collected in June 2003 through Computer Assisted Telephone Interview (CATI) interviews of the students that enrolled at the University of Florence in the year 2001–2002.

The analysis was conducted using regression models which attempted to explain the probability of dropping out by using a set of individual, institutional and contextual variables.

Since the observed response variable is a dichotomous one, the estimation of such probabilities were first made through generalized linear models with classical estimation procedures (maximum likelihood estimation). This was considered as an exploratory phase of the project which would identify a set of significant variables as well as an improved general understanding of the problem. The data were then fitted using a robust approach proposed by Atkinson and Riani (2000), whose results will be reported and commented in another publication.

The procedure we adopted for the estimation of the model classifies all the observations with a hierarchical order in terms of adherence to the model. The characteristics of the extreme groups will be analyzed with descriptive methods and, hopefully, will give us information which could be useful for planning policies that will reduce the university drop out rate.

The use of the robust approach allows identifying singles or groups of students with particular characteristics, for example it may be possible to find groups of individuals who withdrawn the same course programs, otherwise single or groups of freshmen who leaved different course programs. In the first case the explanations should be given to the characteristics of these course programs: it is probable that the learning is too difficult because of the capability and behaviour of instructors, or the organization of classes is poor, or some other reasons due to the teaching activity. The second case, that is groups of withdrawers of different course programs, could depend to specific characteristics of these freshmen (because they are workers, or they got a low score at high school), or even the information about them is biased due to the interviews not correctly carried out or the questionnaire not so clear.

9.3.1.1 Dataset descritption

The analysis on dropout, regarding to all the freshmen enrolled in the past 2001–2002 a.y., used a data set containing some information from the administrative data and some other ones collected from a survey conducted by the Department of Statistics of the University of Florence in June 2003. A number of 2,908 freshmen who left the initial attended course program, which represent the 30% of the total freshmen of that year $(10,053$ cases). From questionnaire the first information we have, is about the different kinds of withdrawal (here called *profiles of drop out*) like moving from one to another course program, or degree program, or even to another university, withdrawal declared with a written communication or not declared.

Then, for all the withdrawers we know:

1. the reasons of changes due to the university activity, such as problems concerning:

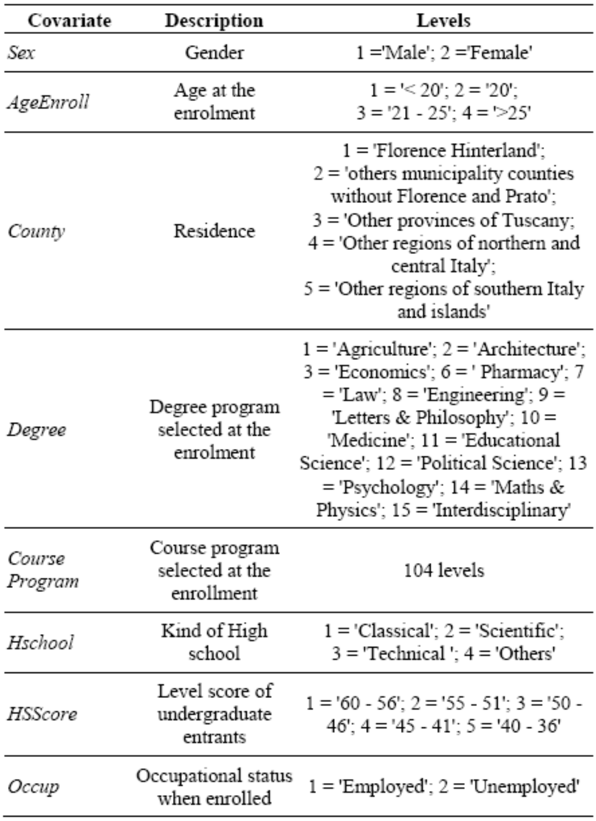
   - the organization of the structures (i.e. if classrooms, laboratories, libraries are adequate to number of students and to technology requested for teaching, etc...);
   - the organization of the course (i.e. amount of class hours with respect to the length of semester, schedules, number of exams during the semesters, etc..);
   - teaching quality of instructors (i.e. clarity, instructor enthusiasm, usefulness of exercises, organization of examinations, the availability of teachers after class, workhome, materials of study, etc...;

2. personal reasons such as: change of residency; health problems; family problems; the occupational status at the enrolment.
   Moreover:
3. in case of moving to another course, degree program or university, which degree and course program they enrolled;
4. in case they did not enroll, or there is no news about their enrolment, whether they intend to enroll again, and eventually in which course/degree program and university.

The response variable of interest in these regression models is the students drop out, identified with the binary variable Y as follows: Y = 1 if dropped out includes all the profiles of withdrawal except to one concerning students who moved to another course or degree program; Y = 0 otherwise.

A preliminary descriptive analysis revealed that among all the variables included in the updated administrative data set, only the covariates shown in Fig. 9.3 yielded a strong association with the response variable.

However, just two covariates are strictly linked to the response variable as indicator of the efficacy and efficiency of teaching activity (i.e. *Course Program* and *Degree Program* selected at the enrolment); whereas the other ones strictly pertain to characteristics of individuals, i.e. gendre (Sex), Age at the enrolment (AgeEnroll), Residence (County), Kind of High school (Hschool), Level score of undergraduate entrants (HSScore), Occupational status when enrolled (Occup).



Covariate	Description	Levels
Sex	Gender	1 ='Male'; 2 ='Female'
AgeEnroll	Age at the enrolment	1 = '< 20'; 2 = '20'; 3 = '21 - 25'; 4 = '>25'
County	Residence	1 = 'Florence Hinterland'; 2 = 'others municipality counties without Florence and Prato'; 3 = 'Other provinces of Tuscany; 4 = 'Other regions of northern and central Italy'; 5 = 'Other regions of southern Italy and islands'
Degree	Degree program selected at the enrolment	1 = 'Agriculture'; 2 = 'Architecture'; 3 = 'Economics'; 6 = ' Pharmacy'; 7 = 'Law'; 8 = 'Engineering'; 9 = 'Letters & Philosophy'; 10 = 'Medicine'; 11 = 'Educational Science'; 12 = 'Political Science'; 13 = 'Psychology'; 14 = 'Maths & Physics'; 15 = 'Interdisciplinary'
Course Program	Course program selected at the enrollment	104 levels
Hschool	Kind of High school	1 = 'Classical'; 2 = 'Scientific'; 3 = 'Technical '; 4 = 'Others'
HSScore	Level score of undergraduate entrants	1 = '60 - 56'; 2 = '55 - 51'; 3 = '50 - 46'; 4 = '45 - 41'; 5 = '40 - 36'
Occup	Occupational status when enrolled	1 = 'Employed'; 2 = 'Unemployed'

Fig. 9.3 Covariates with a stronger association with the response variable.

9.3.1.2 Fitting models

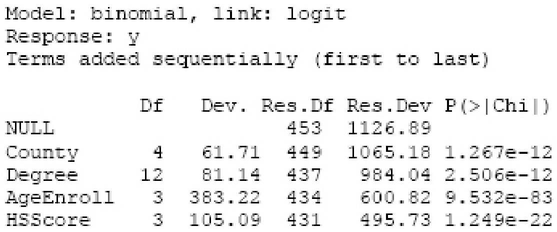
The fit of a logit model for binary data on this reduced data set, allowed to select the significant covariates (AgeEnroll, County, Degree, Hschool, HSScore).

The aim of this study is to detect groups of students, having particular characteristics.

To this purpose the analysis is accomplished by grouping individuals on the basis of the levels of the significant covariates.

The result of this grouping yielded some very low size clusters among 826 groups obtained, so that, as it is known, some test statistics are not reliable anymore. Then, a second fit of a logit model followed using a data set formed by a number of clusters equal to 454, each one at least having more than five units.

The related results led to reject the *kind of high school* covariate (see results in Fig. 9.4).



```
Model: binomial, link: logit
Response: y
Terms added sequentially (first to last)

           Df    Dev. Res.Df Res.Dev P(>|Chi|)
NULL                     453  1126.89
County      4   61.71    449  1065.18 1.267e-12
Degree     12   81.14    437   984.04 2.506e-12
AgeEnroll   3  383.22    434   600.82 9.532e-83
HSScore     3  105.09    431   495.73 1.249e-22
```

Fig. 9.4 Fitting logit model for binomial data.

9.3.1.3 Main results

The algorithm of forward search applied to GLMs, has been carried out with a macro implemented using the R package.

Looking at some forward plots, it is possible to observe the influential importance of some groups.

The first plot to be consider is the goodness of link test along the forward search.

This plot allows to explore different possible link functions (logit, probit, cloglog and arcsin) and then to choice the best one. Looking at the t statistic values of each link put together in a plot (not reported here), it can be noted that trajectories are different and the order of introduction of the observations is different for each link. Moreover, each statistic has an increasing or decreasing trend and it goes outside of bounds (at the 5% level) at different steps of the forward search. Although this means a bad fit in all these cases, from a comparison among trends, it follows that the arcsin link is the most satisfactory one.

Figure 9.5, which reports the goodness of link test of the arcsin function, shows decreasing trend of the statistic as we move towards the end of the forward until it goes out of the significant bounds (5% level) after the step $m = 356$.

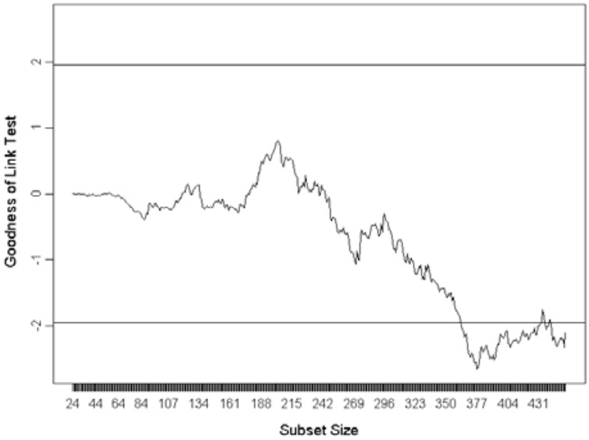


Fig. 9.5 Forward Search: goodness of link tests.

This is due to the presence of groups that differ more than other ones from the bulk of the data; more specifically, they are the last 98 groups entering the subset (as highlighted by the red circle). The presence of observations different from the bulk of data, as well as their affecting the fit of the model, is also highlighted from the monitoring of the deviance of the model.

In Fig. 9.6 it can be noted that the last 98 clusters entering the subset (after the step $m = 356$) cause an exponential increasing of the values of the residual deviance and an exponential decreasing of the values of the Pseudo-R2 statistic; this means that these observations have a significant influence.

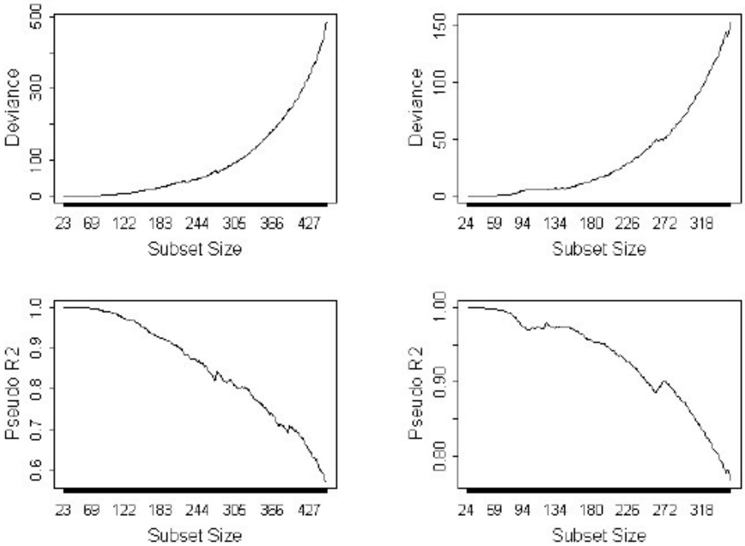


Fig. 9.6 Forward Search: deviance of the Model.

The importance of the effect of the influential groups can be well depicted, once again, by plotting the values of the goodness of link test (5% level). Figure 9.7 reports the three plots of the t statistic during the forward search respectively with the entire data set, after deleting the last 5 and then the last 98 clusters entered the subset. Even though the statistic is inside the bounds after the deletion of the last 98 groups, the trend still shows the bad fit of the model due to all the observations.

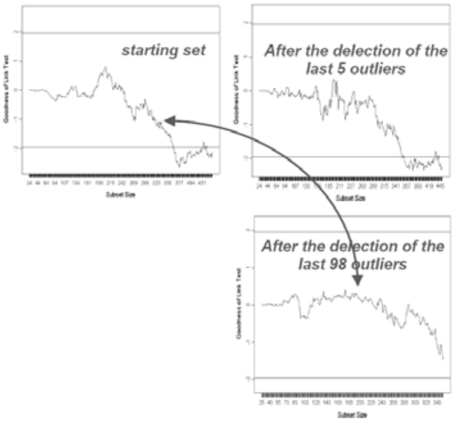
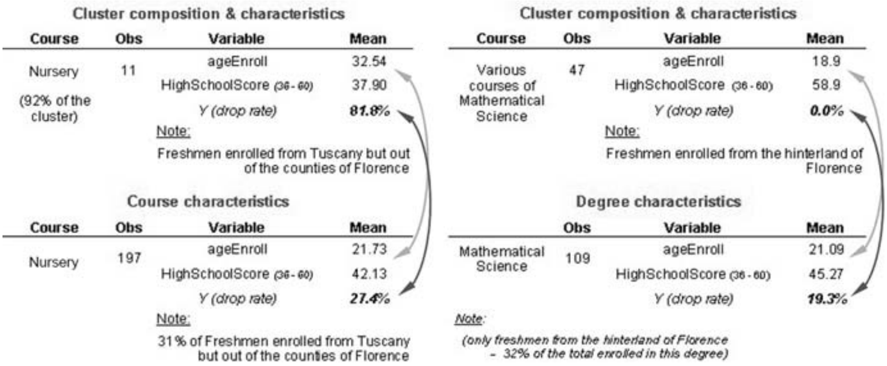


Fig. 9.7 Forward Search: goodness of link tests.

Once the groups of outliers have been detected, the next step should be the investigation of the characteristics of the units inside the groups by the implementation of descriptive analyses, that should allows us to depict various situations useful for the intervention policies aiming to improve the teaching quality and consequently to reduce the drop out rate.
As an example of this kind of analysis, let consider, here, two particular situations arose last steps of the search: the first and the third last groups entering the subset, respectively labeled 270 and 62.



Cluster composition & characteristics

Course	Obs	Variable	Mean
Nursery (92% of the cluster)	11	ageEnroll	32.54
		HighSchoolScore (36 - 60)	37.90
		Y (drop rate)	***81.8%***

Note:
Freshmen enrolled from Tuscany but out of the counties of Florence

Course characteristics

Course	Obs	Variable	Mean
Nursery	197	ageEnroll	21.73
		HighSchoolScore (36 - 60)	42.13
		Y (drop rate)	***27.4%***

Note:
31% of Freshmen enrolled from Tuscany but out of the counties of Florence

Cluster composition & characteristics

Course	Obs	Variable	Mean
Various courses of Mathematical Science	47	ageEnroll	18.9
		HighSchoolScore (36 - 60)	58.9
		Y (drop rate)	***0.0%***

Note:
Freshmen enrolled from the hinterland of Florence

Degree characteristics

	Obs	Variable	Mean
Mathematical Science	109	ageEnroll	21.09
		HighSchoolScore (36 - 60)	45.27
		Y (drop rate)	***19.3%***

Note:
(only freshmen from the hinterland of Florence - 32% of the total enrolled in this degree)

Fig. 9.8 Descriptive analyses of last cluster, number 270 (*left panel*) and of the third last cluster number 62 (*right panel*) entered the subset.

Results reported in Fig. 9.8 show their main characteristics. As concerns the cluster 270, it consists of 11 graduates of nursery, and about all the students (81.8%) dropped out.

It can be noted that with respect to the characteristics of students of the entire nursery course which shows a low drop out rate (27.4%) even including the outlier cluster, they are in average older and less "clever" as highlighted by the lower average of the high school score; moreover, as concerns the residence characteristic they all come from Tuscany region but out of the county of Florence, with respect to the 31% of the entire course program.

This particular situation tells us that the drop out of this course maybe depends more on the characteristics of students than to those of the teaching activity. Anyway, the supplementary information we have from survey, allows us to verify whether this conclusion is correct or some other reasons due to the courses efficiency have to be included.

The second example, indeed, depicts an opposite situation where 47 students attending different courses programs of Mathematical Science do not drop out (0%) and they have in average better performance than the average of all the graduates of the same degree program. About the residence characteristic, all students of the cluster live in Florence and hinterland, while only the 32% of the total enrolled in this degree (also here the number of 109 includes the 47 students of the cluster 62) have the same characteristics.

Here, since the drop out rate of the degree is quite low, we are led to not investigate on the characteristics of these courses but rather on the characteristics of students.

In this particular case, it should have even been interesting to collect, by specific interviews, the evaluations on teaching activities of the courses these outliers attend.

9.4 The Forward Search for ANOVA models

In this section the implementation of the Forward Search method in the ANOVA framework is presented, in order to identify 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 is adapted to the peculiarity of the ANOVA models taking into account the differences between the fixed effects and the random effects ANOVA models. In particular, a procedure is explained to obtain a Robust Forward F Test for the former case and a Robust Forward Likelihood-Ratio Test (*LRT*) for the latter case.

We remind briefly the main characteristics of the one-way ANOVA model. Let y_{ij} be the observed outcome variable of individual i ($i = 1, 2, \ldots, n_j$) within group, or factor level, j ($j = 1, 2, \ldots, J$) where J is the total number of groups and $N = \sum_{j=1}^{J} n_j$ is the total number of individuals. The simplest linear model in this framework is expressed by:

$$y_{ij} = \mu + u_j + e_{ij} = \mu + x_{ij} \tag{9.2}$$

where μ is the grand mean outcome in the population, u_j is the group effect associated with unit j and e_{ij} is the residual error at the lower level of the analysis. This model can be interpreted as a fixed or random effects model, depending on the assumptions about the nature of u_j. When u_j are interpreted as the effects attributable to a finite set of levels of a factor that occur in the data, we have a fixed effect model. On the contrary, when u_j are the effects attributable to a infinite set of levels of a factor of which only a random sample are deemed to occur in the data, we have a random effects model.

Classical assumption on the fixed effects ANOVA model is:

$$e_{ij} \sim N(0, \sigma^2) \ \forall i, j.$$

For the random effects ANOVA model other assumptions are added:

$$\begin{aligned} u_j &\sim N(0, \tau^2) \ \forall j \\ cov(e_{ij}, e_{i'j'}) &= 0 \ \forall i \neq i' \ and \ j \neq j' \\ cov(u_j, u_{j'}) &= 0 \ \forall j \neq j' \\ cov(e_{ij}, u_j) &= 0 \ \forall i, j \end{aligned}$$

and, as a consequence:

$$\begin{aligned} var(y_{ij}) &= var(u_j) + var(e_{ij}) = \tau^2 + \sigma^2 \\ cov(y_{ij}, y_{i'j}) &= \tau^2 \ \forall i \neq i', \end{aligned}$$

where τ^2 expresses the variance among groups and σ^2 expresses the variance within groups.

In the following sections the specific steps to implement the Forward Search are briefly reminded and then an application of the proposed approach to real data, using a set of information referring to the performance of the Italian university system is illustrated. For more details about the theoretical aspects of Forward Search for ANOVA models see Bertaccini and Varriale (2007) and Bertaccini and Varriale (2008).

9.4.1 The Forward Search for the fixed effects ANOVA

9.4.1.1 What is the problem in presence of outliers?

In the fixed effects ANOVA model, we are usually interested in the null hypothesis:

$$H_0 : u_1 = u_2 = ... = u_J = 0,$$

that means that there is not any effect of the factor on the average level of Y.

The statistics used to verify the null hypothesis is defined as:

$$F = \frac{DB/(J-1)}{DW/(N-J)}, \tag{9.3}$$

where DB is the deviance between groups and DW is the deviance within groups. From the normality in the ANOVA assumptions, if the null hypothesis is true, it follows that the ratio statistics F is distributed as a Fisher's F distribution with $(J-1)$ and $(N-J)$ degrees of freedom.

Due to the presence of the sample means in both the DW and DB, the value of the F statistic is strongly affected by the presence of outliers. In fact, it is known that the sample mean is the best unbiased estimator of a population mean under normality assumption, but it shows a strong loss of efficiency in case of contamination or misspecification of the model. This means that in the presence of contaminated data, the "real" value of the first type error probability is systematically higher than the α nominal value (e.g. 0.01, 0.05, ...) and, therefore, the test F will often erroneously reject the null hypothesis.

9.4.1.2 The Forward Search steps

The methodology proposed takes into consideration the presence of groups in the data structure of the ANOVA model. At every step of the Forward Search parameters estimates, residuals, classical F value and other considerable statistics are computed. As usual in the Forward Search method, the procedure is carried on through the classical three steps, that are specified in according to the characteristics of the model:

- Step 1: choice of the initial subset

The specific proposal in the ANOVA framework is to start with the observations y_{ij} that satisfy $min|y_{ij} - med_j|$ in each group j ($j = 1, ..., J$), where med_j is the group j sample median.

- Step 2: adding observation during the search

At each step, the Forward Search algorithm adds to the subset the observations 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 group proportions) of the sample.

- Step 3: monitoring the search

At each stage of the search, parameter estimates, residuals and other relevant statistics, such as classical F test values, are calculated in order to detect the outliers. The main difference between the non-proportional and the proportional approach is that, with the non-proportional strategy the observation belonging to the groups with the minimum variance will enter before the others: hence, the outliers will enter the model last. Instead, in the proportional strategy outliers are forced to enter together with good observations in order to maintain the proportionality of the dimension of the groups.

Finally, a Robust Forward F Test is defined to divide the group of outliers from the other observations, and to evaluate correctly the null hypothesis of fixed effects ANOVA model. The Robust Forward F Test can be defined as a collection $F_{FS} = F(k),...,F(n)$ of the classical F test in each step of the search; 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. The search of the cut-off point can not be "automatic" but is completely based on graphical analysis and is strictly connected to the context of the observed phenomenon.

With the proposed method, the probability of accepting H_1 when H_0 is true is always lower than the same probability obtained with the classical ANOVA F Test.

9.4.2 The Forward Search for the random effects ANOVA

9.4.2.1 What is the problem in presence of outliers?

In a random effects model, such as the ANOVA ones in Eq. (9.2), the observations are aggregated in different levels, so that it is possible to discern first-level units and second-level units or groups. Therefore, we have two different kinds of outliers: first- and second-level outliers. For example, if we consider the hierarchical structure of university system, where students (or first-level units) are aggregated in degree programmes (or second-level units or groups), we could observe one or more students in one or more degree programmes that are anomalous with respect to the student population for some characteristics, or we could observe one or more degree programmes that are anomalous with respect to the degree programmes population. So, we need to focus on the evaluation of the effect of both first and second level outliers on the inferences on the model and, in particular, on their effect on the higher level variance which is statistically evaluated with the LRT.

In many applications of hierarchical analysis, one common research question is whether the variability of the random effects at the group level u_j is significatively equal to 0, namely

$$H_0 : \tau^2 = 0.$$

If the null hypothesis is accepted, then we can conclude that the hierarchical structure of data has no effect on the dependent variable Y. The most used procedure to test this hypothesis is the Likelihood-Ratio Test. In a random effects one-way ANOVA model the asymptotic distribution of the Likelihood-Ratio statistic is a mixture of Chi-squares distributions. However, due to the presence of outliers in the data, the value of the *LRT* statistic can erroneously suggest to reject the null hypothesis H_0 even when there is no second level residual variability.

Therefore, in presence of contaminated data, the classical LRT for the random effects ANOVA model has a similar behavior to the classical F test for the fixed effects ANOVA model. The "true" α value is systematically higher than the nominal ones.

9.4.2.2 The Forward Search steps

The three steps of the Forward Search for the random effects ANOVA model develop in a very similar way to the fixed effects ANOVA model.

- Step 1: choice of the initial subset

As in the previous case, the search starts with the observations y_{ij} that minimize $|y_{ij} - med_j|$ ($j = 1, ..., J$), where med_j is the group j sample median. Moreover, we impose that every group has to be represented by at least two observations; in this way, every group contributes to the estimation of the within random effects.

- Step 2: adding observation during the search

At each step of the search, all the observations are ordered inside each group according to their squared total residuals. The total residuals express the closeness of each unit to the grand mean estimate, making possible the detection of both first and second level outliers. For each group j we choose the first m_j ordered observations and add the one with the smallest squared residual among the remaining.

- Step 3: monitoring the search

At each stage of the search, parameter estimates, residuals and other relevant statistics, such as classical LRT values, are calculated in order to detect the outliers.

Among the most useful outputs there are the plots of the within ($\hat{\sigma}^2$) and between ($\hat{\tau}^2$) variance components and the values of the classical LRT estimated at each step of the Forward Search. See Varriale and Bertaccini (2009) for an analysis of the different trend of the two kinds of plots in presence of first-level outliers or in presence of second-level outliers.

Finally, a Robust Forward LRT is defined to evaluate correctly the null hypothesis of random effects ANOVA model. It can be defined in an analog way to the Robust Forward F test: it is a collection of the values of the classical LR Test statistic computed at each step of the search, and to obtain a Robust Forward LR Test we identify a cut-off point of the progress procedure that best divides the group of observations that differ to the bulk of the data from the others. With the proposed method, the probability of accepting H_1 when H_0 is true is always lower than the same probability obtained with the classical LRT.

9.4.3 The use of the robust ANOVA for the evaluation of the Italian university reform

In this section it is presented an application of the Forward Search for ANOVA models in order to evaluate the impact on the Italian university system of the reform on degree programs, that was enacted in the academic year 2001/02. One of the main aims of this reform was obtain a reduction of the withdrawal rate. The data come from annual surveys conducted by the Italian National University Evaluation Committee (NUEC) during the years 2001, 2002, 2004 and 2005 and refer to the activities of all the public universities during the academic years 1999/2000, 2000/2001, 2002/2003, 2003/2004; the study is limited to the Italian degree programs in Mathematical Science. The dataset is composed by four groups identified by the years in which the NUEC surveys were conducted: two years before the reform and other two after the reform, with 276, 283, 342, 351 observations, respectively. For our purposes, we use the first-year retention rate indicator (*RR*), defined as $RR = 1 - WR$, where WR is the withdrawal rate.

To find out the effect of the reform on the *RT* over different years a fixed effects ANOVA model is estimated, where RT is the dependent variable. First of all, a classical ANOVA F test is conducted: the F value is equal to 2.50 with a p-value equal to 0.058, that is larger than the nominal α value of 0.05. Therefore, on the base of the classical ANOVA F test, the null hypothesis is accepted, and we can conclude that the reform had no effect on the first-year retention rate.

Because of the presence of many outliers in the data set, it is interesting to conduct also a Robust Forward ANOVA F Test, in order to evaluate if the presence of outliers influence the results of the classical test. Among the outputs produced by the Forward Search procedure, the most interesting are shown. In Figs. 9.9, 9.10 and 9.11 are plotted the residual standard error, the estimated RRs, and the classical F values, respectively, at each step of the Forward Search.

The exponential increasing of the curve in Fig. 9.9 confirms the presence of outliers in the dataset, and we can see that they enter the model during the last steps of the Forward Search. From Fig. 9.10 it can be seen that the estimated RRs referring to the two years after the reform (2004 and 2005) are almost systematically higher than the two others. Only when the outliers enter the model the F values for years 2004 and 2005 converge to the F values referred to years 2001 and 2002. Finally, form the analysis of Fig. 9.11 is evident that the F statistics is always in the reject region: only when the outliers enter in the procedure, the F values fall in the acceptance region.

Therefore, on the basis of the results shown in Figs. 9.10 and 9.11 it can be concluded that: (i) by taking into account only data conformed with the ANOVA model hypotheses, the university system reform had a positive effect on the increasing of the first year retention rate; (ii) the presence of outliers induces to an opposite, and wrong, conclusion. This example highlights the superiority of the Forward Search approach in comparison with a classical approach, such as classical ANOVA F test, for the inference in presence of outliers.

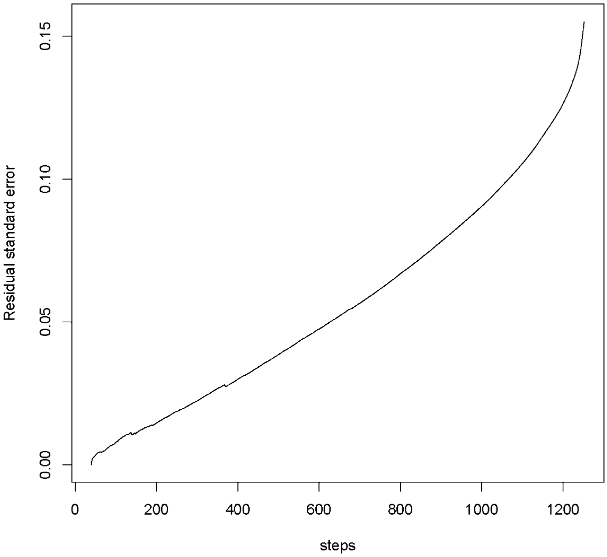


Fig. 9.9 Forward plots of the residual standard error.

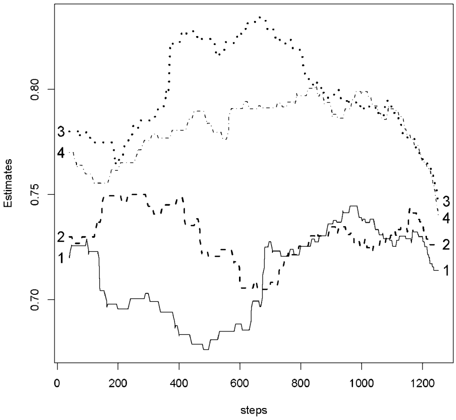


Fig. 9.10 Forward plots of the estimated coefficients for the first year retention rate of the Italian degree programs in Mathematical Science.

9.5 Concluding remarks

The peculiarities of *Forward Search* and its analytical possibilities, which have been highlighted in this work, make it an approach that is usually preferred to other robust methods; in fact:

1. *Forward Search* combines robustness and efficiency because, during the evolution of the analysis, the estimation procedures are based on well-known statistical algorithms (maximum likelihood, least squares,...) with proven efficiency

and quick computation abilities; in other words, no ad hoc high intensity computation algorithms are required for estimating parameters;

2. The approach is easily extended to different analytical contexts (regression, generalised linear models, multivariate method of analysis, etc.) and is therefore applicable to most of the situations of which multidimensional data are available;
3. The method can be generalised even to cases in which there is auto-correlation between the observations (historical sets, models for spatial data);
4. The approach features a higher degree of generality compared to other robust methods, since the *outliers* are neither deleted nor “underweighted”; *Forward Search* actually allows their entrance probably in the final stages of the procedure, thus offering the analyst the possibility of evaluating the effects on the inferential conclusions drawn from the adapted statistical model.
5. Finally, we know that the analyses and the representation of a certain degree of the performance of university are a useful support for planning some interventions and actions as concern the organization of the structures, but especially the teaching activities. To perform a deeper studies about the complex system of relationships and factors which affect problem like for example the drop out of University, or the impact of a new reform, it is necessary to use appropriate analytical models. The robust diagnostics regression analyses are able not only to supply an answer to these needs, but also allow identifying observations or groups of units (outliers) having specific characteristics. The inspections on these outliers could really help to implement university programmes on the teaching activities aimed at improving the quality of this service.

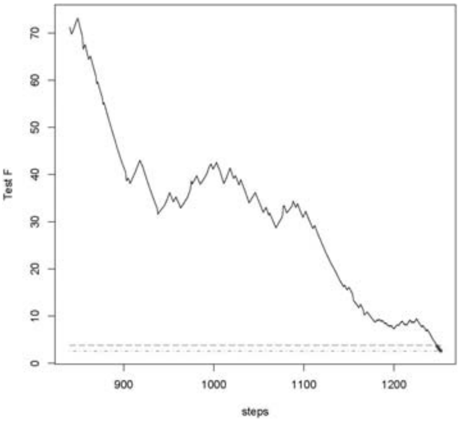


Fig. 9.11 Forward plots of the F statistics.