Evaluating the University Educational Process. A Robust Approach to the Drop-out Problem

Matilde Bini, Bruno Bertaccini

Statistics Department "G. Parenti, University of Florence, Italy

Summary. The use of robust procedures in regression model estimation identifies outlier data that can inform on specific subpopulations. The aim of this study is to analyse the problem of first year dropouts at the University of Florence. A set of administrative data, collected at the moment of enrolment, combined with the information gathered through a specific survey of the students enrolled in the 2001-2002 academic year at the same athenaeum, was used for the purpose. In order to identify the most important variables affecting the students' dropout, the data were first fitted with generalized linear models estimated with classical methods. The same models were then estimated with robust methods that allowed the detection of groups of outliers. These in turn were analysed to determine the personal or contextual characteristics. These results may be relevant for the implementation of academic policy changes.

Keywords: Dropout rate; Outliers; Forward search method.

1. Introduction

The evaluation of the higher education system – and particularly of the university one – and the use of statistical methods for measuring its performance have become important issues given the cultural, social and economic relevance of this educational level. The statistical analyses carried out on the Italian university system highlight, among the other aspects, its weaknesses (Bini, 1999; Bertaccini, 2000; Chiandotto & Bertaccini, 2003).

We argue that the average levels of single indicators do not emphasize the 'net' effects of factors that quantify the indicators, because of the existing interactions among these and grouping variables. Hence, it is necessary to implement appropriate analytical models for representing the relationships with the phenomenon considered.

The linear regression model satisfies this requirement. The applications described in the following are the basis for a complete and complex analysis performed using the *Forward Search* algorithm (Atkinson & Riani, 2000). This method may be used to make correct decisions both for the allocation of resources and verify the planned objectives of the educational programmes. In fact, this method is able also to reveal the presence of unusual characteristics of the phenomenon under study.

The procedure starts by fitting the model with a number of observations sufficient for estimating parameters and continues the fitting of the model to increasing subsets. The units are ordered according to their proximity to the fitted model. If the model agrees with the data, the robust and least squared procedures yield similar parameters and error estimates. However, these may change considerably with the *Forward Search*. The monitoring of the changes and of some statistics used to make inference in regression models allows reaping information useful not only for detecting outliers, but also, and above all, to comprehend their importance in inference making.

We present an application of detection of observations with characteristics that can explain the low degree of withdrawal from university programmes where the drop-out problem is particularly significant.

In Section 2 we introduce the estimation problems caused by the presence of outliers, in Section 3 we show the properties of the least median squares as the robust approach to the regression model; Section 4 is devoted to the presentation of the forward algorithm applied to classic linear models and to generalized linear models. Concluding remarks are presented in the Section 5.

2. The problem of outliers

In various fields of research, the regression model is a common statistical tool. The properties of the Ordinary Least Squares (OLS) estimators justify its popularity but not the mistreatment that occasionally occurs with their use, when insufficient attention is given to both verification of the specification theories and the presence of anomalies in the data at hand.

It is to be remembered that the estimate of the p parameters in a regression model depends on p statistics computed on the whole dataset; if any of these differ from the bulk of the data, the fitting process can conceal these differences or, otherwise, be strongly influenced by them.

The outliers can derive from mistakes performed during the recording steps, or from unusual phenomena, or can identify units accidentally included in the sample but belonging to other populations.

The response variable is not the only factor that can undergo irregularity. Outliers may be related with explanatory variables because of the larger frequency with which atypical data may be collected.

If the regression parameters were known, there would be no difficulty in detecting the outliers as units enhancing the highest terms of error. However, difficulty arises when the parameters of the model have to be estimated with observations that may contain abnormal units.

To sidestep their presence, *robust* methods of estimation (so called because they can produce estimates that are not influenced by contaminated data) have been proposed.

These methods identify as *outliers* the units that reveal very high residuals[1]. Methods for investigating several outliers simultaneously entail the use of robust techniques for organising observations according to their residuals. One that deserves special mention is the *Forward Search* algorithm (Section 3).

3. The *Forward Search* for classical linear models

The ordinary least squares method yields estimates of β coefficients that minimise the residual sum of squares. The distance $(y_i - \hat{y}_i)^2$ gives higher weights to units with higher residuals. Therefore, if there are few observations with very large residuals, ε_i, the estimates are strongly affected, particularly those concerned with high leverage points. Because of this sensitivity, estimation methods that yield good estimates in the presence of contaminated data are termed "*robust*" (Box & Andersen, 1955).

Donoho & Huber (1983) introduced the *breakdown point* (Rousseuw, 1987; Atkinson & Riani, 2000) that is the smallest fraction of contamination that can induce a certain estimator *T(Z')* to assume values arbitrarily distant from *T(Z),* where $Z = (X, y)$ stands for the matrix $n \times (p + 1)$ of the original data and Z' for all the possible contaminated samples.

Contaminated are the samples obtained by substituting any m number of the original observations with arbitrary values. No reference is made in this definition to the probability distribution of the data.

Among the various robust approaches, special mention must be made to the studies by Huber and to the *Least Median of Squares* (LMS) estimation method (Rousseuw, 1984). This latter estimator has a 50% breakdown point, i.e. at least half of the observations must be outliers in order to have repercussions on the estimates. Even if the contamination rate is lower, the method

[1] An alternative approach is the *diagnostic analysis*, which implies the computation of statistics that detect abnormalities and the data most liable to exert an influence (Cook & Weisberg, 1982; Atkinson, 1985). The main drawback of these procedures is that, as the number of potential outliers increases, there is a combined explosion of the possible subsets under investigation. In robust regression, the model is adapted with techniques that do justice to the bulk of the data and examine the units that mostly differ from the predicted values. Nevertheless, robust approaches and diagnostic analyses often produce the same results.

gives unbiased estimates of the regression hyperplane, provided n is large. This is the maximum *breakdown point* a regression model can tolerate[2].

The *Forward Search* algorithm (Atkinson & Riani, 2000) combines the diagnostic ability in identifying groups of *outliers* with the properties of the robust estimation methods, particularly the LMS. The basic steps of this algorithm are (Section 3.1): choice of the best starting-set free of outliers, the addition of observations, and the monitoring of the statistics that detect outliers.

3.1 The choice of the initial subset

The best initial subset is detected with the least median of squares approximation (Rousseuw, 1984), which guarantees a starting-set free of outliers.

The basic steps of the algorithm are the following. Let $Z=(X,y)$ be a $n \times (p+1)$ matrix.

If n is not too large and $p<<n$, the choice of the initial subset derives from the enumeration of the $\binom{n}{p}$ distinct p-tuples $S^{(p)}_{i_1,\ldots,i_p} \equiv \{z_{i_1},\ldots,z_{i_p}\}$, where $z^T_{i_j}$ is the i_j-th row of Z $(j=1,\ldots,p)$ and $1 \le i_j \ne i_{j^*} \le n$.

Let $i^T=[i_1,\ldots,i_p]$ and $e_{i,S^{(p)}_\imath}$ be the least squared residual for the i-th unit since the regression model is fitted only with the observations in $S^{(p)}_\imath$. Hence, the starting-set is the set $S^{(p)}_*$ of p-tuples that satisfies

$$e^2_{[med],S^{(p)}_*} = \min_\imath \left[e^2_{[med],S^{(p)}_\imath} \right], \tag{1}$$

where $e^2_{[k],S^{(p)}_\imath}$ is the k-th ordered square residual among $e^2_{i,S^{(p)}_\imath}$, $(i=1,\ldots,n)$ and *med* is the integer part of $(n+p+1)/2$.

If $\binom{n}{p}$ is large, the subset is chosen by using (1) and enumerating 3,000 p-tuples of matrix Z.

Therefore, the *Forward Search*, according to the size of the data set and to the capacity of computers, finds the minimum of (1) among the 3,000 p-tuples

[2] The very robust behaviour of the LMS estimator is in contrast with the Ordinary Least Squares (OLS). In OLS, a sensitive variation in the estimations can take place when just one outlier assumes arbitrary large values; since the *1/n* fraction tends towards zero as the sample size increases, the breakdown point will be zero.

from a sample of size n. If the number of p-tuples is less than 3,000, the *Forward Search* will take account of all the subsets.

3.2 The addition of observations

Let $S_*^{(m)}$ be a subset of size $m \geq p$; the *Forward Search* procedure moves towards the subset $S_*^{(m+1)}$ by selecting units that have the first $m+1$ ordered residuals $e^2_{[k],S(m)}$. The procedure ends when all the observations are included in the subset; that is, when $S_*^{(m)} = S^{(n)}$. Therefore, the *Forward Search* estimator $\hat{\beta}_{FS}$ is the set of the $n-p+1$ ordinary least squared estimators obtained at each step of the procedure; that is

$$\hat{\beta}_{FS} = \left(\hat{\beta}^*_p, \hat{\beta}^*_{p+1}, \ldots, \hat{\beta}^*_{n-1}, \hat{\beta}^*_n \equiv \hat{\beta}_n\right).$$

The change of the dimension from m to m+1 implies in general the entrance of only one unit into the previous subset. Nevertheless, two or more units could enter $S_*^{(m+1)}$ while one or more could leave.

This last situation may occur, though less frequently, when an observation that belongs to a group of outliers emerges during a study. In fact, in the subsequent step, the remaining units of the group of outliers reveal a less typical behaviour and a few can enter the subset at the same step.

This new approach combines the features of the least median squares with the efficiency of the OLS estimators. The robustness of the approach consists in the continuous inclusion of units in the subset free of outliers, rather than in the choice of a particular estimator having a high breakdown point. In other words, this method is not particularly affected by the technique used to select the initial subset of units, since it generates a subset free of outliers, or with masked outliers that may be removed at following steps of the search.

3.3 The monitoring of the procedure

The estimate of σ^2 changes during the procedure since, at each step of the search procedure, m units with the smallest residuals enter the subset ($m=p+1, \ldots, n$). Therefore, even without outliers, we have $s^2_{S(m)} \leq s^2_{S(n)}$ given $m < n$. Usually, the curve drawn by $s^2_{S(m)}$ shows an initial slight increase, as it happens when the data and the adopted model agree, and a steep increase if significant outliers are present.

A very important graphical representation is the one that allows checking the behaviour of all *n* residuals at each step of the *Forward Search*. Large values of residuals from units not included in the data point out the presence of outliers.

Since $s^2_{S^{(m)}_*}$ is strongly dependent on *m*, all residuals are standardized with the mean of the squared residuals s^2. The plot that shows the trajectories of leverage values is useful for detecting outliers. At any step, a unit enters the subset $S^{(m)}_*$, and corresponding leverage $h_{i,S^{(m)}_*} = x_i^T \left(X^T_{S^{(m)}_*} X_{S^{(m)}_*} \right)^{-1} x_i$ values are plotted.

At start, the search of the subset $S^{(p)}_*$ includes only *p* observations, each one with a leverage value equals to one. Thereafter, these values decrease. The outliers that enter the $S^{(m)}_*$ subset at the final steps of the analysis may show higher leverage values than the others, though it is possible that units that make up the initial set show the highest values for the whole analysis.

The *Forward Search* algorithm can be extended to Generalized Linear Models (GLMs) that are an extension of linear regression models where the response variable is not normally distributed, so that the expected values are modelled with a link function (Agresti, 2002). In particular, if models for binomial data are adopted, as in this work, the search algorithm is the same as the linear regression one, except that squared residuals of deviance d_i^2 are used rather than least square residuals. In this case, the procedure starts with a random selection of subsets in *p* dimension and chooses the subset with the smallest value of the median of the deviance[3].

The afore-mentioned situations highlight that *Forward Search* algorithm can be effective to the sample by combining its diagnostic capability in identifying groups of outliers together with the properties expressed by robust methods of estimation. The results obtained when the algorithm is applied are twofold: anomalous units can be eliminated to produce a model that is more stable and conform to reality, and, on the other hand, the information resulting from the analysis of the structural composition of the identified outliers can be used to perform further investigation into the phenomenon under study.

In this respect, the application was successful in evaluating both the effectiveness of university titles in relation to employment and the discontinuation

[3] This is the general rule except in the case of models for binary data. In fact, except for the response variable, the number of zeros is not equal to number of ones, the Least Median of Squares method of estimation will include in the initial subset only observations with frequent responses. Hence, in order to keep both types of response in equilibrium during the procedure, the search algorithm must be modified.

of university studies, and proper for supplying indications that may be helpful for improving the quality of the educational process.

4. An analysis of the dropout rate in Florence

There has been a significant evolution in the Italian public administration over the recent years. In particular, the new regulations acknowledge the Ministry of Education, University and Research as the responsible entity for defining the targets and general strategies for developing the tertiary educational system and its assessment, while the universities are granted ample self-governance, even though part of the funds are subject to the fulfilment of specific requisites[4].

The decentralization and self-government, and the restriction on funding, imply that the universities, as responsible for the results achieved, must necessarily perform procedures of intense and exhaustive self-assessment in terms of both efficiency and internal and *external effectiveness*.

One of the most important indicators of internal efficiency is the dropout rate, which continues to override 50% in many universities, even after the recent reform of the study cycles, with serious repercussions on programming university policies and even on the society as a whole.

Florence University data show that during the 2002/03 academic year almost 30% of the enrolled students dropped out of the course they chose in the previous year. Because of the severe consequences this phenomenon has on educational programming, Florence University carried out a *CATI (Computer Aided Telephone Interviewing)* survey on the possible causes of dropout[5] of the students enrolled in the 2002/2003 academic year and who resulted to be in one of the following conditions:

- *Passed* (P) to another degree course in Florence University;
- *Transferred* (T) to another university;
- *Renounced* (R) explicitly to the studies;
- *Implicit* (I) if he/she is not enrolled in 2002/03 and does not belong to any of the previous categories;
- *Suspended* (S), e.g., for national military service; these students could also be included in the Implicit category.

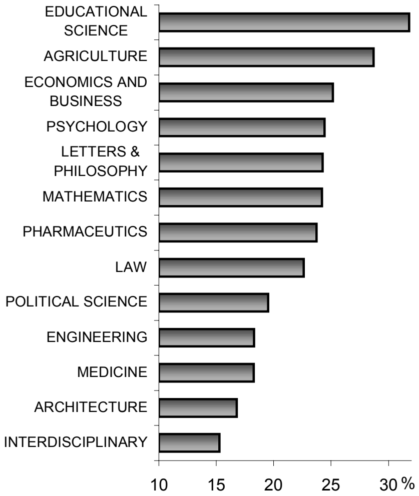
The complete database includes a student's profile (gender, age, residence), high school curriculum (type and final marks), faculty and study programme upon first registration, employment status and position regarding compulsory

[4] Financial, managerial and organisational self-government of every Italian university was introduced under Laws no. 168/89, 537/1993, 59/1997 and 127/1997.

[5] The survey was realised thanks to funds of *CAMPUS*ONE and *OUTCOMES* projects. The data collected permitted the correction (due to errors or delayed input by the Student secretariat) of some data in the university files.

Table 1. Distribution of the dropout rate of the sample by faculty

Degree programme	*Response Y*	
	Enrolled	**% (*Y*=1)**
AGRICULTURE	*358*	28,7
ARCHITECTURE	*1,065*	16,8
ECONOMICS AND BUSINESS	*1,029*	25,2
PHARMACEUTICS	*197*	23,8
LAW	*777*	22,6
ENGINEERING	*889*	18,3
LETTERS & PHILOSOPHY	*1,518*	24,3
MEDICINE	*717*	18,3
PSYCHOLOGY	*1,429*	24,4
EDUCATIONAL SCIENCE	*462*	31,8
MATHEMATICS	*649*	24,2
POLITICAL SCIENCE	*450*	19,5
INTERDISCIPLINARY	*513*	15,3
TOTAL	***10,053***	22,3



military service at the time of enrolment, as well as the correct form of registration and drop-out motives of transfer to another programme within Florence University or to another university.

The analysis was performed from the following perspectives: dropouts can be considered a phenomenon depending exclusively on the organisation of the study programme or on the central organisational policies of the university. The transfer from one course to another may be an important rupture in a student's career since it involves a waste of time and resources both for the student and, indirectly, for the study programme. However, the number of students who change programme within the same university is attributable to the sudden awareness of the incompatibility between their personal aptitudes and the subject matter of the programme.

The response variable is:

Y=1, if the student has dropped out of the initial enrolment programme and did not pass to any other programme at Florence University the year after (the options in the questionnaire were *transferred*, *renounced*, *implicit* or *suspended*);

Y=0, if the student is still registered in the original programme, or has transferred or submitted a request to transfer to another study programme within Florence University.

Preliminary analyses indicated the association of some covariates with the response variable[6] (Table 1). Only the faculty and the study programme re-

[6] The contingent of students reduced to 9007 because of the omissions in the administration files corresponding with the variables related to the phenomenon under study; nevertheless, this did not alter the distribution of drop-out rates shown in Table 1.

Table 2. List of the explanatory variables correlated with response variable

Factor	Description	Levels
DEGREE	Degree programme selected on enrolment	1 = AGRICULTURE 2 = ARCHITECTURE 3 = ECONOMICS AND BUSINESS 6 = PHARMACEUTICS 7 = LAW 8 = ENGINEERING 9 = LETTERS & PHILOSOPHY 10 = MEDICINE 11 = EDUCATIONAL SCIENCE 12 = POLITICAL SCIENCE 13 = PSYCHOLOGY 14 = MATHEMATICS 15 = INTERDISCIPLINARY
COURSE	Course programme	104 levels
GENDER	Gender	1 = Male; 2 = Female
COUNTY	Area of residence	1 = Florence - Hinterland 2 = Other towns in provinces of Florence & Prato 3 = Other provinces in Tuscany 4 = Other northern and central regions 5 = Southern regions and Islands
HSCHOOL	Kind of high school	1 = Classics; 2 = Scientific 3 = Technical; 4 = Others
HSSCORE	Final marks	1 = 60 – 56 2 = 55 - 51 3 = 50 – 46 4 = 45 - 41 5 = 40 – 36
AGEENROLL	Age at enrolment	1 = Under 20 2 = 20 3 = 21 – 25 4 = over 25
WORK	Occup. status at enrolment	1 = Employed 2 = Unemployed

sulted to be related to the quality of the organisation of the education, whereas the others related to the profile of the students (Table 2).

All the variables correlated with the phenomenon under study were considered as possible predictors[7] in a logistic model for binary data, automated with a *backward elimination* option. The model is merely exploratory, since it is used to detect the subsets of explicative variables with the highest powers of discrimination. In Table 3, we illustrate the main results.

For the best possible assessment of the impact of education organisation as a possible cause of drop-out, the *Forward Search* algorithm was applied to the "global data"[8] in order to develop a logistic model for binomial data. In this

[7] The variables associated with the original faculty were deleted from the list of possible explanatory variables. In addition, the numeric-type variables, such as average marks at exams and upon degree, were categorised. The loss of information connected with this transformation caused negative consequences that are nevertheless insignificant for our purpose, but are advantageous from a computational and interpretative point of view.

[8] The term global data means that all observations showing the same levels of the investigated factors are grouped together.

case, the unusual data will be groups of subjects with a common profile. Analysis of the structural composition of these groups might reveal, for example, contingencies sufficiently significant to justify the moderate dropout rates within contexts where the dropout phenomenon is particularly serious.

Table 3. Main results of the logistic model for data grouped according to the levels of the factors, identified at the start of the analysis

```
Summary of glm(formula = y ~ County + Degree + AgeEnroll + HSscore,
  Family = binomial(link = "logit"), data = abnd03, weights = enr)

  Deviance Residuals:
         Min          1Q      Median          3Q         Max
  -2.8730178  -0.8601809   0.0091846   0.7433140   2.9480225

  Coefficients:   Estimate Std. Error z value  Pr(>|z|)
  (Intercept)   -1.298283   0.155187 -8.3659 < 2.2e-16 ***
  County2        0.256007   0.081815  3.1291 0.0017534 **
  County3        0.342038   0.073441  4.6573 3.203e-06 ***
  County4        0.654752   0.108170  6.0530 1.422e-09 ***
  County5        0.708523   0.098155  7.2184 5.260e-13 ***
  Degree2       -0.593890   0.167855 -3.5381 0.0004030 ***
  Degree4       -0.189403   0.275305 -0.6880 0.4914699
  Degree5       -0.115360   0.171284 -0.6735 0.5006290
  Degree6       -0.164708   0.170176 -0.9679 0.3331091
  Degree7       -0.177111   0.155681 -1.1376 0.2552667
  Degree9       -0.603253   0.179289 -3.3647 0.0007663 ***
  Degree10       0.143121   0.174118  0.8220 0.4110916
  Degree11      -0.342498   0.200154 -1.7112 0.0870500 .
  Degree49      -0.295448   0.156798 -1.8843 0.0595299 .
  Degree52      -0.045730   0.163929 -0.2790 0.7802749
  Degree56      -0.274399   0.187682 -1.4620 0.1437303
  Degree58      -0.606601   0.204628 -2.9644 0.0030326 **
  AgeEnroll2     0.552379   0.078484  7.0381 1.949e-12 ***
  AgeEnroll3     0.958635   0.079691 12.0293 < 2.2e-16 ***
  AgeEnroll4     1.262348   0.114303 11.0438 < 2.2e-16 ***
  HSscore2      -0.374010   0.063411 -5.8982 3.676e-09 ***
  HSscore3      -0.710940   0.097891 -7.2626 3.797e-13 ***
  HSscore4      -0.983368   0.110800 -8.8752 < 2.2e-16 ***

  Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of Fisher Scoring iterations: 3
Null deviance: 1126.89  on 453  degrees of freedom
Residual deviance:  495.73  on 431  degrees of freedom
AIC: 1688.75
  Analysis of Deviance Table
  Terms added sequentially (first to last)
               Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
  NULL                           453   1126.889
  County        4   61.712       449   1065.178 1.2666e-12
  Degree       12   81.135       437    984.042 2.5057e-12
  AgeEnroll     3  383.222       434    600.820 9.5323e-83
  HSscore       3  105.091       431    495.730 1.2492e-22
```

The subjects are grouped according to the covariates detected during the preliminary stage (*faculty, residence, age at enrolment* and *final marks of high school*). It is well known that if the binomial denominators are not large enough, the usual goodness of fit statistics do not meet with any known distribution. Moreover, it is likely that very small groups, which yield very little information, might disturb the search algorithm. Hence, we decided to remove groups containing less than five individuals.

The model for binomial data adapted to the new dataset is shown in Table 3. The groups thus created are 454. The deviance of this model demonstrates that adaptation is more than sufficient (the residual deviance is just slightly higher than the levels of free residuals). The effects of the factors and the single levels are evident, with the exception of those referring to certain faculties that do not show significant effects on the response due to the high levels of the standard errors. Once the working model has been defined, the *Forward Search* algorithm is applied.

An important result is the graph of the statistical test for the goodness of fit of the link function (Figure 1). The test values demonstrate a decrease beyond the limits of significance[9] in the final part of the search algorithm, due to the presence of groups that differ considerably from the bulk of the data.

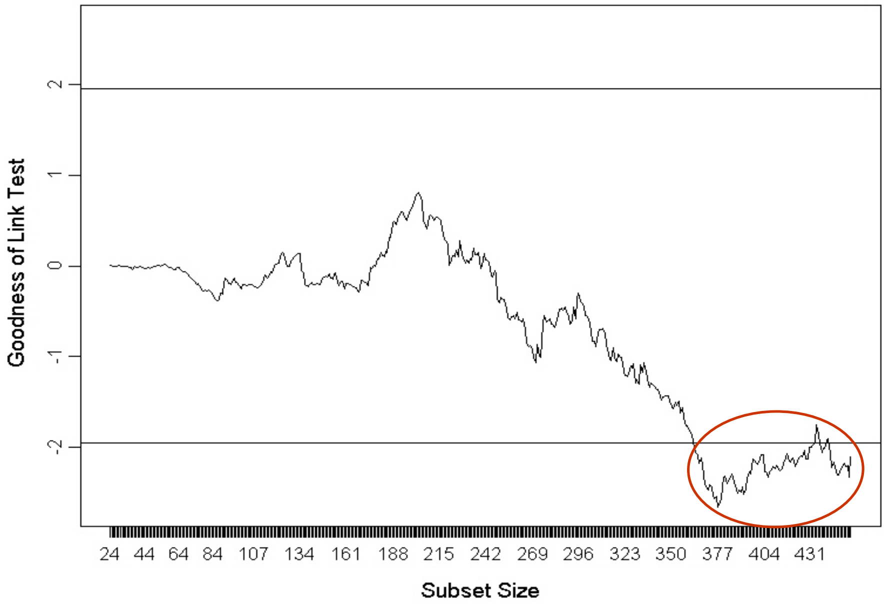


Figure 1. Forward Search: goodness of fit test of the link function

[9] The test has been structured so that values statistically equal to zero (1 - α = 95%) tend towards the goodness of the function chosen.

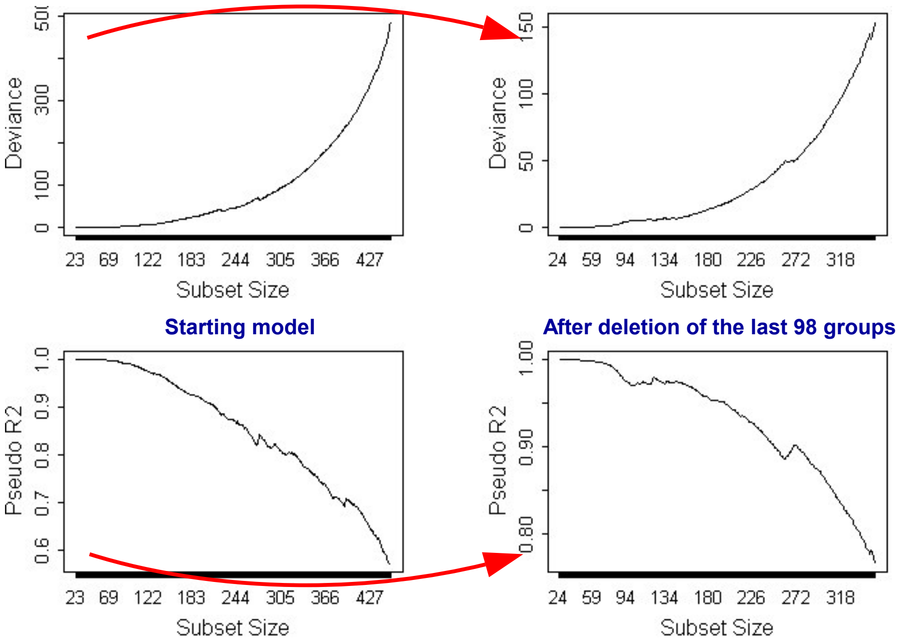


Figure 2. Forward Search: model deviance before and after the deletion of groups of outliers

Therefore, the last 98 groups are worthy of further investigation. The presence of observations with abnormal trends can be detected also from the graphs referring to explained deviation and R^2 index (Figure 2). The last 98 groups cause an exponential increase in the residual deviation (from 150 to 500) and a great decrease in the Pseudo R^2 values (from 0.80 to 0.60).

It is interesting to see how the results would have been different after deletion of the groups at the final stages of the search. The deletion contributed evidently to the adaptability and stability of the goodness of fit test (Figure 3).

As an example, we can examine the composition of group 270, the last one that entered in the analysis. It is composed of 12 subjects attending Medicine, 11 of which are enrolled in the Nursing programme. The dropout rate is 83.3% of students (10 out of 12 students) and 81.8% with reference to the programme (9 out of 11); this latter value is very high in comparison with the mean of the overall course enrolments (27.4%, see Table 4).

This may depend on the profile of the group: these are subjects much older at the enrolment and with lower high school marks than the others. Moreover, they lived in Tuscany but not in the provinces of Florence and Prato, whereas the percentage of students residing in these provinces for the entire programme duration was 27.4%. This suggests that dropout is attributable more to personal characteristics of the students than to the organisation of the educational programme.

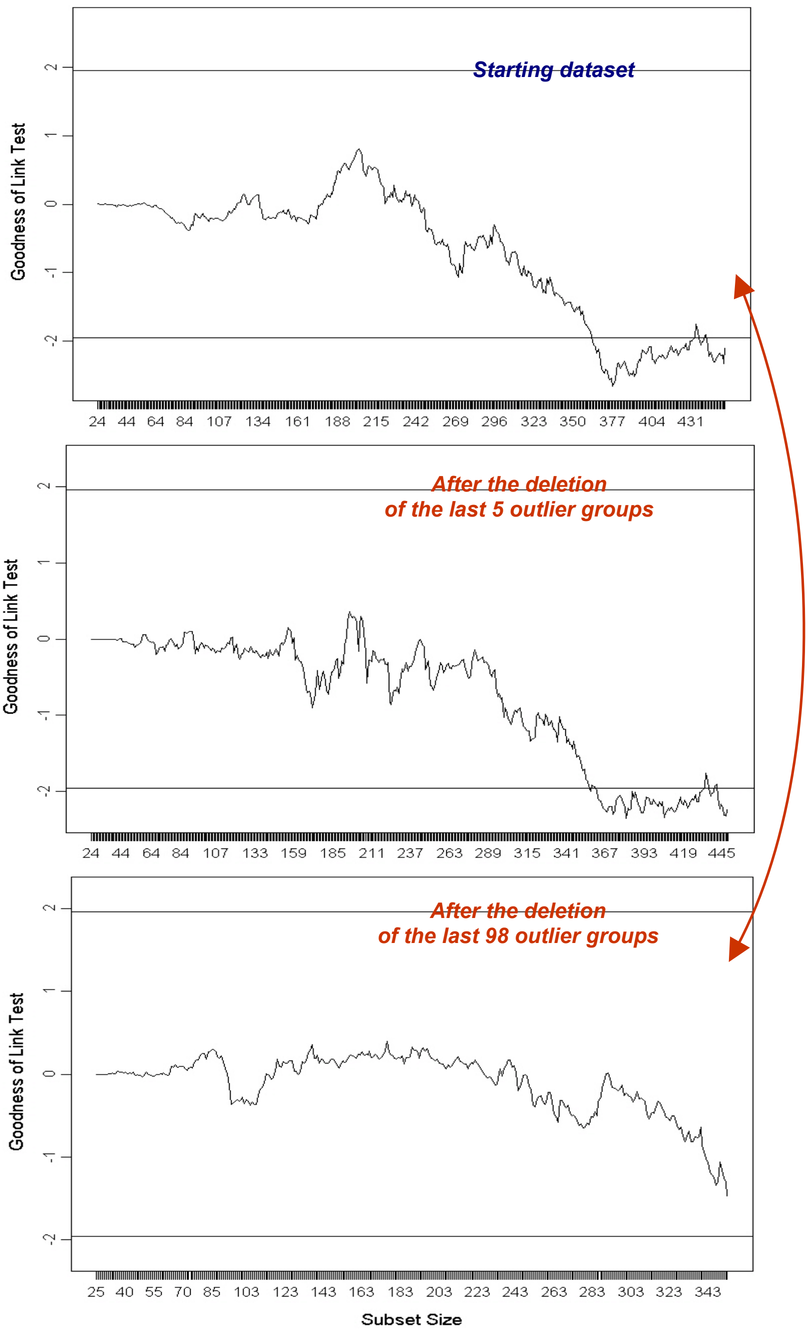


Figure 3. Forward Search: goodness of fit test of the link function before and after the deletion of groups of outliers

Table 4. Analysis of the composition of cluster 270

Cluster composition and characteristics

Course	Obs	Variable	Mean
Nursing	11	AgeEnrollment	32.54
		HighSchoolScore (36-60)	37.90
(92% of the cluster)		*Y (drop rate)*	***81,8%***

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

Course characteristics

Course	Obs	Variable	Mean
Nursing	197	AgeEnrollment	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 and Prato

5. Concluding remarks

The analyses performed, together with the evidence of a certain level of performance in university education, are useful tools for programming and organising facilities and educational activities.

Nonetheless, in order to carry out a more thorough analysis of the complex system of relationships and factors that influence the drop-out rate, it will be necessary to use specific analytical models. The robust diagnostic tools applied to regression analysis are capable not only of supplying a reply to this necessity, like the regression models based on traditional estimation methods, but can also identifying the units or groups of units with particular characteristics. These observations may be a source of information that is useful for defining new educational programmes, for improving their quality and, as a consequence, for reducing the rate of drop-out at university.

References

AGRESTI A. (2002) *Categorical Data Analysis, 2nd Ed.*, Wiley, New York.

ATKINSON A. C. (1985) *Plots, Transformations and Regression*, Oxford University Press, Oxford.

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

BERTACCINI B. (2000) *Misure di efficacia esterna dell'istruzione universitaria: indicatori statistici e analisi robusta* (B.A. dissertation), University of Florence.

BINI M. (1999) *Valutazione della Efficacia dell'Istruzione Universitaria rispetto al Mercato del Lavoro*. Rdr 03/99. Osservatorio per la Valutazione del Sistema Universitario - Ministero dell'Università e della Ricerca Scientifica e Tecnologica.

CHATTERJEE S., HADI A. S. (1988) *Sensitivity Analysis in Linear Regression*, Wiley, New York.

CHIANDOTTO B., BERTACCINI B. (2003) *Profilo e sbocchi occupazionali dei laureati e diplomati dell'Ateneo fiorentino nell'anno 1999*, University of Florence.

COOK R. D., WEISBERG S. (1982) *Residual and Influence in Regression*, Chapman and Hall, London.

DONOHO D.L., HUBER P.J. (1983) The notation of breakdown point, In: BICKEL P.J., DOKSUM K., HODGES J.L.Jr (eds) *A festschrift for Erich Lehmann*, Wadsworth Inc., Belmont, CA: 157-184.

ROUSSEEUW P.J. (1984) Least Median of Square Regression, *Journal of the American Statistical Association,* ***85***: 633-639.

ROUSSEEUW P.J., LEROY A.M. (1987) *Robust Regression and Outlier Detection*, Wiley, New York.