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.