Grazyna Dehnel (Poznań University of Economics and Business)
Elżbieta Gołata (Poznań University of Economics and Business)

Robust regression in the monthly business survey

Topic

Keywords: business statistics, robust regression, short-term statistics, M-estimation, outliers

Introduction

Market economy generate growing demand for information at low level of aggregation provided by businesses on regular basis, at short intervals.

One of the main problems connected with estimating population parameters across domains are distributions of auxiliary variables. This is due to, among other factors, the presence of outliers. In many business surveys sample sizes are large enough to compensate for the presence of outliers, which have a relatively small impact on estimates. However, at low levels of aggregation, the impact of outliers might be significant. Therefore, in the case of a population of enterprises the classical approach should be accompanied by technics that are resistant to the occurrence of outliers. One of the more important is robust regression. It comprises a number of methods (M-estimation S-estimation LTS-estimation, MM-estimation). Until recently the implementation of them was limited owing to their iterative nature. With advances in computing power and the growing availability of statistical packages such as R, SAS or Stata the applicability of robust regression methods has increased considerably.

Choosing the robust regression method it is necessary to make additional decisions about its parameters and functions. The study was limited to an analysis of the properties of one of these methods – M-estimation.

Methods / Problem statement

The paper was aimed at evaluating properties of M-estimators, using data from a survey of small and medium-sized businesses in the transport section of the classification of economic Activities (NACE Rev.2). The comparison involves nine M-estimators, each based on a different weighting function. The results and conclusions are formulated on the basis of empirical data from the DG-1 business survey.

The analysis was divided into two parts. The first part involved assessing the quality of the model's goodness of fit based on the robust version of the coefficient of determination and estimation errors of the equation parameters. Differences in values obtained for each type of M-estimator reflect their sensitivity to the presence of different kinds of outliers (in the x-direction or in the y-direction) and their distance from the bulk of the data. Analysis of the results suggests that the use of M-estimation improves the goodness of fit of the model only when y-outliers are present. In the case of x-outliers, the application of M-estimation resulted in lower values of the coefficient of determination (compared to OLS).

The second stage of the analysis consisted in comparing the properties of parameter estimators for the regression equations derived on the basis of the weighting functions. The bootstrap method was applied to determine measures for the assessment of the efficiency, bias and MSE. 1000 iterations of drawing samples were made, which were then used to calculate: relative estimation error (REE), mean absolute relative bias (ARB), relative root mean square error (RMSE).

Results / Proposed solution

The first stage of the analysis involved assessing the distribution of businesses in terms of variables included in the model. In addition, Student's t-test and Cook's D confirmed the presence of outliers. These properties indicated the need for the use of robust regression method.

The analysis was divided into two parts. The first part involved assessing the quality of the model's goodness of fit based on the robust version of the coefficient of determination and estimation errors of the equation parameters. Differences in values obtained for each type of M-estimator reflect their sensitivity to the presence of different kinds of outliers (in the x-direction or in the y-direction) and their distance from the bulk of the data. Analysis of the results suggests that the use of M-estimation improves the goodness of fit of the model only when y-outliers are present. In the case of x-outliers, the application of M-estimation resulted in lower values of the coefficient of determination (compared to OLS).

The second stage of the analysis consisted in comparing the properties of parameter estimators for the regression equations derived on the basis of the weighting functions. The bootstrap method was applied to determine measures for the assessment of the efficiency, bias and MSE. 1000 iterations of drawing samples were made, which were then used to calculate: relative estimation error (REE), mean absolute relative bias (ARB), relative root mean square error (RMSE).

Conclusions

• The use of the M-estimator in the presence of outliers can considerably improve the quality of the model's fit compared to the classical method of estimation – it largely depends on the type of outliers. The M-estimator is only resistant to y-outliers but is not resistant to leverage points. It should therefore be used in situations where there are no leverage points.

• In practical applications of M-estimation, the selection of function is not a key choice for obtaining good robust estimates. The adoption of each of the nine weighting functions analyzed in the study yielded similar results from the viewpoint of values of estimated parameters and their standard errors. The least adequately fitted models were those based on Cauchy's and Hampel's functions. The best fit was obtained for the models based on Fair's and Huber's functions; one drawback in their case was the relatively high level of standard errors.

• The largest gain in efficiency and robustness of M-estimators was obtained when Talworth's and Tukey's functions were used. This result was particularly visible for domains in which the influence of outliers on the quality of the classical LS model was very strong. Owing to the curve shapes of Talworth's and Tukey's functions, observations with large residuals are ignored.