Elements of Robust Regression for Data with Absolute and Relative Information

Karel Hron[1] and Peter Filzmoser[2]

Robust regression methods have advantages over classical least-squares (LS) regression if the strict model assumptions used for LS regression are violated. We briefly review LMS and LTS regression as robust alternatives to LS regression, and illustrate their advantages. Furthermore, it is demonstrated how robust regression can be used if the response variable contains relative rather than absolute information.

Keywords: multiple linear regression, robustness, relative and absolute information, compositional data

1 Introduction

In multiple linear regression we consider a linear combination of several explanatory variables, and use this aggregated information to predict a response variable. It results in estimations of parameters of a linear functional that reveal how the response depends on the set of explanatory variables. The least-squares method that is commonly used to obtain the estimations, leads to the best statistical efficiency if certain model assumptions are fulfilled. On the other hand, this method is also very sensitive to outlying observations that could completely destroy the results and thus make any interpretation meaningless. For this reason, many robust counterparts were proposed in the literature. They are usually less efficient than the classical approach, but they are in general substantially more resistant to outliers or other deviations from the underlying regression model assumptions. The robust methods thus represent a practical and meaningful alternative to the classical approach, as far

Palacký University, tř. 17. listopadu 12, 77146 Olomouc, Czech Republic hronk@seznam.cz · Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria P.Filzmoser@tuwien.ac.at

as both the response variable and the covariates carry absolute information. However, in many areas data occur which include only relative information (known nowadays under the term *compositional data*) where all the relevant information is contained in the ratios rather then in the absolute values as in the usual case. As these data induce another sample space, they need to be transformed before regression analysis is carried out.

This contribution is organized as follows. In Section 2 a brief review of the classical and robust regression estimators is provided. In Section 3 the basic concepts of compositional data are presented. The final section shows how the relative information can be used in (robust) regression analysis using a real data example.

2 Classical and Robust Linear Regression

Multiple regression analysis forms a tool for prediction of values of a quantity, the response variable, using known (independent) variables. The main task is to find a functional relationship (here assumed to be a linear one) between the response and covariates, i.e. to estimate parameters of the regression function [4]. Let $x_1, \ldots, x_q$ be the q variables that we use for prediction of the response variable y. Under the standard regression assumptions, y is a random variable and $x_1, \ldots, x_q$ are assumed to be non-random. Having n observations of both y and the explanatory quantities, the linear multiple regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \varepsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{1}$$

or in matrix form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

with the n-dimensional vector $\boldsymbol{y}$ containing the observations of the response variable, the random vector of errors $\boldsymbol{\varepsilon}$ (are assumed to have mean zero), and the $n \times (q+1)$ dimensional design matrix $\mathbf{X}$ with full column rank. Under the assumption of uncorrelated components ε_i, with variance $\mathrm{var}(\varepsilon_i) = \sigma^2$, the vector of unknown parameters can be estimated using the least-squares (LS) method as

$$\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{3}$$

Obviously, the estimate $\widehat{\boldsymbol{\beta}}_{LS}$ minimizes the term

$$\sum_{i=1}^{n} \varepsilon_i^2(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}. \tag{4}$$

It is easy to verify that $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, and under the additional assumption of normality of $\boldsymbol{\varepsilon}$ it is also the maximum-likelihood estimator of $\boldsymbol{\beta}$. Consequently, it can be used to obtain the *predicted values*

$\widehat{\boldsymbol{y}}_{LS}$ of $\boldsymbol{y}$ as

$$\widehat{\boldsymbol{y}}_{LS} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LS} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is known as the *hat matrix*. The *estimated residuals* are

$$\widehat{\boldsymbol{\varepsilon}}_{LS} = \sqrt{\sum_{i=1}^{n} \varepsilon_i^2(\widehat{\boldsymbol{\beta}}_{LS})} = \boldsymbol{y} - \widehat{\boldsymbol{y}}_{LS} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \tag{6}$$

where $\boldsymbol{I}$ stands for identity matrix of order n.

LS-estimation may fail if the model assumptions are violated. Data points deviating from the linear trend can have a strong influence on the estimation because LS regression is based on the squares of the residuals, which then can become very large. We now illustrate this effect in linear regression with one predictor variable.

Figure 1 (left) shows five points that approximately follow a linear trend. Moving one observation in y-direction has a strong influence on the LS parameters, because also the regression line follows this movement in y-direction (right). Also the robust regression method LTS regression (see below) has been applied here, and the movement of the point has no effect on this estimate: the dashed line representing the resulting LTS line coincides with the LS-line of the original data (dotted).

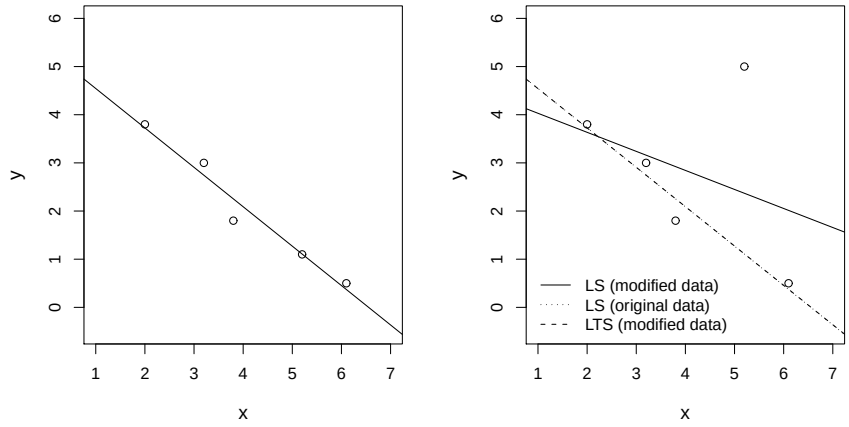

Fig. 1 Influence of an outlier in y-direction on classical LS and robust LTS regression.

An even worse behaviour is shown in Figure 2, where in the left picture a similar design is presented as in Figure 1 (left). When now an observation is moved in x-direction, the LS regression line is completely changed (right). For this reason, x-outliers are also called *leverage points* because they can "lever" the LS regression line. This undesirable behaviour of LS regression can be avoided by robust regression. The solution of LTS regression for the

modified data is almost the same as that for LS regression for the original data.

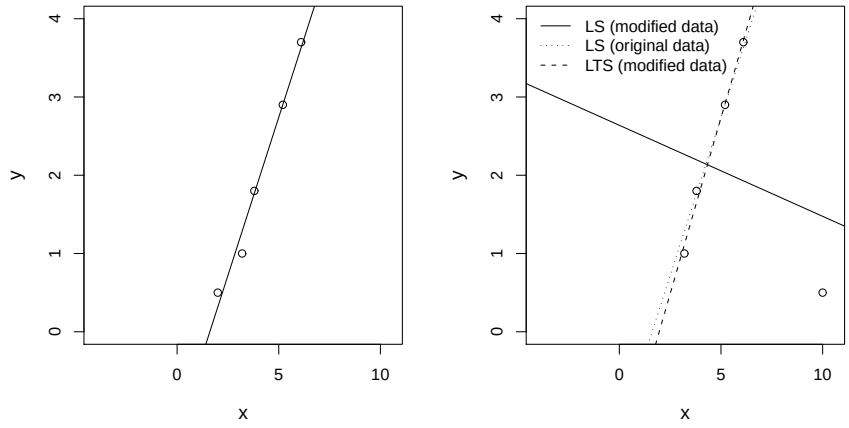

Fig. 2 Influence of an outlier in x-direction on classical LS and robust LTS regression.

The basic principle of robust regression is to fit the model to the data majority that follows the linear trend [5]. Accordingly, for *Least Median of Squares* (LMS) regression the function

$$\text{median}_i \, \varepsilon_i^2(\boldsymbol{\beta}) \tag{7}$$

is minimized. Here, the sum from (4) is simply replaced by a median. However, any explicit solution for the regression coefficients as for LS regression is not available, it has to be found using approximative algorithms. For LMS regression it turns out that up to 50% of the data points can be moved arbitrarily without any substantial change of the regression coefficients. This behaviour is expressed by the *breakdown point* which equals 0.5.

Another very robust regression method is *Least Trimmed Sum of Squares* (LTS) regression, where the term

$$\sum_{i=1}^{h} (\varepsilon_i^2(\boldsymbol{\beta}))_{(i)} \tag{8}$$

is minimized, again using a numerical procedure. Here $(\varepsilon_i^2(\boldsymbol{\beta}))_{(1)} \leq \cdots \leq (\varepsilon_i^2(\boldsymbol{\beta}))_{(n)}$ are the sorted squared residuals. By taking $h \approx n/2$, the method has a breakdown point of about 0.5, for larger h it moves to $(n-h)/n$.

3 Relative Information and Compositional Data

As far as the response variable y carries absolute information, the preceding considerations can be used directly. However, in many practical situations the information is not absolute but relative, often expressed in proportions or percentages. Examples of relative information are the unemployment rate in selected countries, proportions of people working in agriculture, percentages of inhabitants with tertiary education, or proportions of the household budget spent on foodstuff. Here the usual model assumptions fail because the values of the response variable are bounded in a certain interval, e.g. in $(0, 100)$ in case of percentages, and the assumption of normal distribution is thus not meaningful. However, the problem is in fact a conceptual one and it is inherent to the nature of the data. Namely, here the idea of the relative scale is quite an intuitive concept of differences for them. While the difference between 5% and 10% is the same as between 45% and 50%, the proportions show a quite contrasting relation, because 5% is half of 10%, while 45% is 0.9 of 50%. Thinking in terms of differences in ratios is natural for this kind of data, called in general compositional data (or compositions for short) [1], where only the relative information is of interest. They induce the simplex as the sample space with an own geometry, called nowadays the Aitchison geometry. Thus, compositional data need to be moved from the simplex to the usual Euclidean space isometrically before any statistical analysis can be carried out. This causes in fact that the relative information is transformed into absolute information. The best transformation for this purpose seems to be the isometric logratio (ilr) transformation [2], for both theoretical and practical reasons.

Here we consider a situation where only the response variable includes relative information, but not the explanatory variables. Thus we deal with the problem of an univariate analysis of compositional data [3]. In this case, the ilr transformation of the response variable y simplifies to a new variable (that reminds to the well-known logit transformation)

$$z = \frac{1}{\sqrt{2}} \ln \frac{y}{c - y} \ , \tag{9}$$

where c corresponds to the total value of the whole (1, 100%, total amount of inhabitants working in agriculture, total household budget in Euro) for each observation. After ilr transformation, the values can already be used for regression analysis in the sense of the previous section. After regression analysis and a corresponding prediction for z, the results can be back-transformed to obtain an interpretation in the sense of the original variable y.

4 Use of Robust Regression for Compositional Data

To demonstrate the theoretical considerations numerically, we apply regression analysis to an example where the relation between the percentage of employees in the tertiary sector and the value of the Gross Domestic Product (GDP) per capita in the member states of the European Union is investigated. The considered data are from the year 2009. The tertiary sector is also called "service" sector, where service provision is defined as an economic activity that does not result in ownership, and this is in contrast to providing physical goods. The GDP is a basic measure of a country's overall economic output. It is the market value of all final goods and services made within the borders of a country in a year. The data were obtained from public sources of the internet encyclopedia Wikipedia. Figure 3 (left) shows the data without Luxembourg, where the response variable is already ilr-transformed. Thus both variables contain absolute information and the regression analysis in sense of the previous section can be applied. In the lower right corner of the plot an outlier is clearly visible: Ireland, with a GDP of 30.900 Euro per capita, but with only 49% of employees in tertiary sector. This outlier can be considered as y-outlier, because it is still not exceptional in x-direction. Still, a strong effect on LS estimation (solid line) is visible, LTS regression is not affected by the outlier, and when excluding Ireland from the analysis, LS would practically coincide with the LTS line.

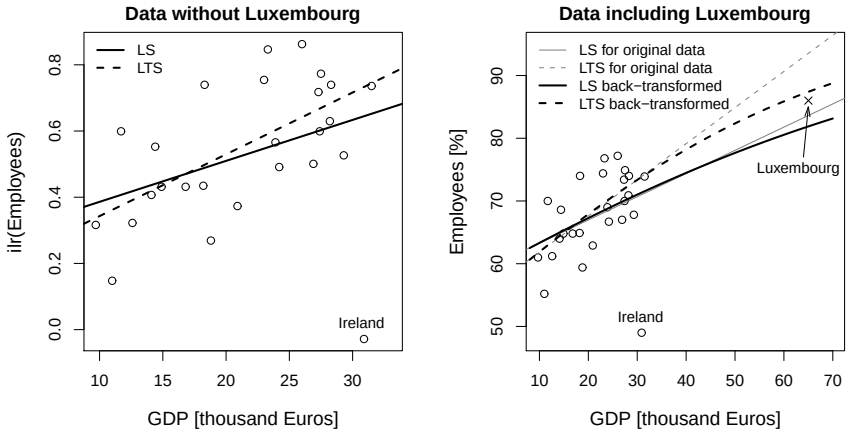

Fig. 3 Regression analysis for the percentage of employees in the tertiary sector after ilr transformation (response variable) and the GDP per capita (explanatory variable) in the member states of the European Union.

Figure 3 (right) shows the original data, together with the regression lines from the left picture back-transformed to the original space. Note that the back-transformation is unique, because from Equation (9) we obtain y by

$$y = \frac{c \cdot \exp(\sqrt{2}z)}{1 + \exp(\sqrt{2}z)} \ . \tag{10}$$

Due to the different geometry of the simplex, the back-transformed regression lines are no longer linear. To make the effect of the ilr transformation visible, classical and robust regression is also applied in the wrong geometry using the original y-variable. The results are shown by gray lines. Now Luxembourg is projected into the plot. The GDP of Luxembourg is exceptionally high with 65.009 Euro per capita, and 86% of the employees are in the tertiary sector. The prediction from LTS regression in the ilr-space is closest to the true value, while LS regression, as well as regression analysis (classical and robust) applied in the wrong geometry differ substantially. The reasonability of the robust approach applied in the ilr-space is confirmed by the fact that the resulting regression line is almost unchanged if the outlier Luxembourg is included already at the beginning of the analysis.

References

1. Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London
2. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Math Geol 35:279–300
3. Filzmoser P, Hron K, Reimann C (2009) STOTEM 407:6100–6108
4. Johnson R, Wichern D (2007) Applied multivariate statistical analysis (6th edn). Prentice-Hall, London
5. Maronna R, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, New York