PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Krzysztof Jajuga
Wroclaw University of Economics

SOME MODELS OF UNIVARIATE AND MULTIVARIATE DATA IN STATISTICAL ANALYSIS

1. Introduction – types of statistical analysis

The development of the use of the statistical methods in real world applications has become very substantial in recent years. There are many factors of this development, including: the progress in the area of computer technology, the growth of the number of available data and the increasing need of practice to analyze data. On the other hand, statistical and econometric theory offers a great number of different data analysis methods.

The aim of this paper is to present some models to be used in statistical data analysis. However, at the beginning we concentrate on some general discussion how the different data analysis models can be systematized. From the point of view of the link between theory and practice, it is clear that almost any type of statistical analysis can be viewed through the following general equation:

DATA = MODEL + ERROR.

Therefore one has to understand and explain the available data by imposing some kind of model. Since, in most real life applications, model is only an approximation of the reality, the inclusion of error is needed.

Traditionally, one can distinguish between two types of models:

- confirmatory models – where the model is built to verify an underlying hypothesis provided by theory;
- exploratory models – where the model is built to search for some "patterns" in data and these "patterns" possibly can lead to some hypotheses.

Clearly, from the point of view of real applications, confirmatory models are better; however, when the hypothesis is not available or it is rejected, then one has to try exploratory models.

Our classification of types of statistical analysis is done with respect to four different criteria:

1. Randomness of the variables used in the model.

Here, clearly, one can distinguish:

- stochastic approach, where model contains random variables, so the study of statistical distributions is crucial one;
- descriptive approach, where the variables are not random, therefore model is understood in the purely approximation sense.

2. Homogeneity of data set.

Here, we can distinguish:

- unified approach, where data set is considered as a homogeneous one, therefore the parameters characterize the whole data set;
- clustering approach, where data set is considered as containing several (or more) classes, and the characteristics (parameters) are different for each class.

3. Concentration on center or outliers of data set.

Here, we can distinguish:

- classical analysis, where one concentrates on typical characteristics of data set, describing "core" of this set;
- tail analysis, where one concentrates on the non-typical characteristics of data set, describing tails of the distribution (outlying observations).

4. Type of considered data.

Here, we have two types of analysis:

- time series analysis,
- cross-sectional data analysis.

The main difference between these two types of analysis occurs in the stochastic approach, since for time series in a general case each observation may come from separate population.

In the next sections we describe several different models used for data analysis. These models are very often suitable in many practical situations.

2. Simplest models in univariate data analysis

By no doubt, the simplest type of statistical analysis is univariate analysis, conducted in descriptive approach as core analysis. Here the basic characteristics of data set, that are considered, are:

- location parameter;
- scale parameter;
- skewness parameter;
- kurtosis parameter.

Now we present a possible framework to analyze these parameters. This framework is based on L_p-norm, very often used in classification and data analysis.

Location parameter

In L_p-norm, location parameter is the solution to the problem of the minimization of the following function, subject to the sought location parameter μ:

$$\left(\sum_{i=1}^{n}|x_i-\mu|^p\right)^{1/p} \tag{1}$$

Clearly, the function given by (1), can be understood as a measure of overall goodness of fit of location parameter to set of univariate observations. This criterion is expressed through the appropriate distances of data from the location parameter.

By solving the problem of minimization of (1) we get the following "natural" parameters:

- in L_1-norm: median;
- in L_2-norm: arithmetic mean;
- in L_∞-norm: midrange, being the average of the maximal and minimal value of all observations.

Scale parameter

First of all, let us notice, that scale parameter is a location-dependent parameter. This leads to the conclusion that scale parameter should measure the "volume" of the set of observations, possibly through a scatter of observations around the location parameter. Assuming still L_p-norm framework, we define scale parameter as:

$$\sigma=\frac{1}{n^{1/p}}\left(\sum_{i=1}^{n}|x_i-\mu|^p\right)^{1/p} \tag{2}$$

By assuming different values of p, we get the following scale parameters:

- in L_1-norm – mean deviation from the median:

$$\sigma=\frac{1}{n}\sum_{i=1}^{n}|x_i-\mu|, \tag{3}$$

- in L_2-norm – standard deviation:

$$\sigma=\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(x_i-\mu)^2}, \tag{4}$$

- in $L\infty$-norm – half of range:

$$\sigma=\frac{1}{2}(x_{\max}-x_{\min}). \tag{5}$$

Skewness parameter

Similarly as scale parameter, skewness parameter is location dependent. It can be regarded as a parameter comparing a "volume" of the observations below and above location parameter. Assuming still L_p-norm framework, we define skewness parameter as:

$$sk = \frac{\left(\sum\limits_{x_i > \mu} |x_i - \mu|^p\right) - \left(\sum\limits_{x_i < \mu} |x_i - \mu|^p\right)}{\left(\sum\limits_{i=1}^{n} |x_i - \mu|^p\right)}. \tag{6}$$

The denominator of (6) can be considered as a "volume" of the data set. This "volume" can be split into two parts: "volume" of the part of observations above the location parameter and "volume" of the part of observations below the location parameter. The numerator of (6) compares these two parts by taking the difference between them. Clearly, the proposed parameter is normalized to lie in the interval [-1;1]. The closer this parameter to 0, the closer to the symmetric is the considered data set. Also, there is natural interpretation of signs, positive means skewness to the right, negative means skewness to the left. There is one drawback of this proposal, since it does not work for $L\infty$-norm – it takes always value to zero.

Kurtosis parameter

Similarly as scale and skewness parameters, kurtosis parameter is also location dependent. It can be regarded as a parameter comparing a "volume" of the observations close to the location parameter and the observations lying in the tail. Assuming still L_p-norm framework, we propose to define kurtosis parameter as:

$$kt = \frac{\left(\sum\limits_{d_i > l} |x_i - \mu|^p\right) - \left(\sum\limits_{d_i < l} |x_i - \mu|^p\right)}{\left(\sum\limits_{i=1}^{n} |x_i - \mu|^p\right)}. \tag{7}$$

The denominator of (7) can be considered as a "volume" of the data set. This "volume" can be split into two parts: "volume" of the part of observations close to the location parameter and "volume" of the part of observations far from the location parameter. Here appropriate threshold distance, denoted by l, should be taken into account. The numerator of (7) compares these two parts by taking the difference between them. Clearly, the proposed parameter is normalized to lie in the interval [0;1]. The closer this parameter to 0, the higher kurtosis is observed in the considered data set.

3. Some other models of univariate data analysis

The presented simple type of univariate analysis can be extended or modified in several ways. One of the most straightforward ways of the generalization is to adapt clustering approach, in which we consider data set as heterogeneous one, containing several classes. Then we have as many location parameters as classes.

Still assuming L_p-norm framework, we define location parameters in clustering approach as the solution to the minimization problem of the following function:

$$\left(\sum_{j=1}^{K}\left(\sum_{i\in C_j}\left|x_i-\mu_j\right|^p\right)\right). \tag{8}$$

The function (8) can be interpreted as a generalization of function (1), by introducing distinct location parameter for each class, given that the number of classes, denoted by K, is known. Clearly, we get the same solution as for simple problem, where the respective location parameter is determined by using the observations belonging to the particular class.

In practice, the classification of observations is not known, therefore the iterative algorithm is applied, where in the consecutive iterations the location parameters are determined using given classification, and then the classification is updated, by assigning the observation to the class with the nearest location parameter. Of course, one can find the other parameters (scale, skewness, kurtosis) for each class, in similar way as in the unified (homogeneous) approach.

All the presented models were based on the descriptive approach. Now we move to the stochastic approach.

In this approach we can distinguish two possible models for univariate data:

- statistical distribution;
- stochastic process.

The first model is usually used when one analyzes cross-sectional data and it is reasonable to assume that all the observations come from the same population. The second model is usually used when analyzes time series data and the natural assumption is to deal with series of random variables. In both cases such models are used, where stochastic parameters are related to location and scale parameters.

Although stochastic process is the most general framework, one often analyzes conditional distribution in time moment given (observed) previous values of the process; this is the main approach in financial econometrics and dynamic econometrics. The general model for univariate time series analysis is given as (e.g. [Tsay 2002]):

$$X_t = g(F_{t-1}) + \sqrt{h(F_{t-1})}\varepsilon_t, \tag{9}$$

$$g(F_{t-1}) = \mu_t = E(X_t \mid X_{t-1},....), \tag{10}$$

$$h(F_{t-1}) = \sigma_t^2 = V(X_t \mid X_{t-1},....). \tag{11}$$

In (9) time series is decomposed into two parts. The first part, given by (10), is simply the expected value of conditional distribution and the second part, given by (11), is the variance of conditional distribution. So in the model two parameters are considered, location parameter and scale parameter. However, they are defined for conditional distribution of random variable corresponding to considered time period, given random variables corresponding to previous periods.

Of course, the model given in (9)-(11) could be specified by assuming different types of functions g and h, as well as the distribution of random component, denoted by ε.

The last univariate type of analysis, we present here, is tail analysis, conducted in the stochastic approach. There are two main models in tail analysis, sometimes referred to as the parts of Extreme Value Theory. These are the following models:

- distribution of maximum or minimum;
- conditional excess distribution

As far as the first model is concerned, one analyzes the function of random variables defined as a maximum (or minimum) of a set of random variables. It is known that the limiting distribution of the maximum belongs to the family of the so-called Extreme Value Distributions (containing Fréchet distribution, Gumbel distribution and Weibull distribution).

In the second model one analyzes the distribution given through the following density function:

$$f(X|X > u). \tag{12}$$

So this is conditional distribution of the considered variable given that it takes value above some threshold – it is the analysis in the right tail (for left tail we take values below some threshold). It is known that this distribution belongs to the class of Generalized Pareto Distributions. Both models of Extreme Value Theory are well described by Embrechts, Klüppelberg and Mikosch [1997].

4. On some models of multivariate data analysis

By no doubt, the models of multivariate data analysis are more complex than those suitable for univariate analysis. The main reason is that multivariate analysis is not just simple "multiplication" of univariate analysis. The additional "dimension" of the analysis is the dependence between the components of a set of variables. We present here some general remarks on multivariate models.

In the descriptive approach, the commonly used models refer to some characteristics of data set, as in the univariate case. The possible models, very often studied in theory and used in applications, are:

- location vector (including mean vector);
- scatter matrix (including covariance matrix);
- regression function;
- principal components;
- canonical variates.

In the stochastic approach, two basic models are:

- multivariate distribution (for cross-sectional data);
- multivariate stochastic process (for time series).

It is worth to mention, that in the multivariate approach one assumes very often that the considered distribution belongs to the class of elliptically symmetric distributions. There the stochastic parameters are linked to the descriptive parameters: location vector and scatter matrix. The latter one contains scale parameters and dependence parameters.

The alternative approach that can be used in multivariate analysis is copula analysis. This approach allows to separate the univariate properties from the dependence structure of multivariate data.

The idea of copula analysis lies in the decomposition of the multivariate distribution into two components. The first component consists of the marginal distributions. The second component – the crucial one – is the function linking these marginal distributions in multivariate distribution. This function reflects the structure of the relationship between the components of the multivariate random vector. Therefore the analysis of multivariate distribution function is done by „separating" univariate distribution from the relationship.

This idea is reflected in the so-called Sklar theorem [Sklar 1959], given through the following formula:

$$F(x_1,\dots,x_n)=C(F_1(x_1),\dots,F_n(x_n)), \tag{13}$$

where: F – the multivariate distribution function;

F_i – the distribution function of the i-th marginal distribution;

C – copula function.

Thus the multivariate distribution function is the function of the univariate (marginal) distribution functions. This function is called copula function and it reflects the structure of the relationships between the univariate components.

References

Embrechts P., Klüppelberg C., Mikosch T. (1997), *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.

Sklar A. (1959), *Fonctions de repartition à n dimensions et leurs marges*, Publications de l'Institut de Statistique de l'Université de Paris, 8, pp. 229-231.

Tsay R.S. (2002), *Analysis of Financial Time Series*, Wiley, New York.

WYBRANE MODELE DANYCH JEDNOWYMIAROWYCH I WIELOWYMIAROWYCH W ANALIZIE STATYSTYCZNEJ

Streszczenie

Omówiono niektóre modele danych, będące podstawą analizy statystycznej, osobno dla danych jednowymiarowych i wielowymiarowych. W przypadku danych

jednowymiarowych są to dwie podstawowe grupy modeli: parametry położenia, skali i inne parametry charakteryzujące zbiór obserwacji, rozkład jednowymiarowy. Z kolei w przypadku danych wielowymiarowych są to następujące podstawowe grupy modeli: wektor parametrów położenia, macierz parametrów rozrzutu, parametry zależności, rozkład wielowymiarowy, kształt zbioru punktów w przestrzeni wielowymiarowej. W każdej grupie przedstawione i porównane są możliwe sposoby określania konkretnych modeli. Rozważania teoretyczne zilustrowane są badaniami empirycznymi. Opracowanie ma charakter uogólniający i porządkujący wiele różnorodnych podejść w klasycznej analizie statystycznej.