PROBABILISIC REGRESSION STRUCTURES

Agnieszka Szumera

SUMMARY

Classical problem of regression deals with determining the polynomials W of the degree not greater than a given number $n \in \mathbb{N}$, which are optimally fitted to given finite sequences x and y of numbers representing empirical data. As an optimization criterion the method of least squares is adopted here. Polynomial regression play an important role in the analysis of numerical data. The simplest type of polynomials regression are so called linear regressions, formed by polynomials of the degree not greater than one. Linear relationship is also very intuitive. That is the reason why linear regressions are most often used in practice. On the other hand, this approach is not precise enough in certain cases which touch more complex relationship.

The aim of the research is to develop a mathematically consistent and possibly the general theory of regression used in probability theory, statistics and econometrics, where determination of the best fitted theoretical model to empirical data is based on the concept of the least squares method. The dissertation introduces a new concept of *probabilistic regression structure* $\mathfrak{P} := (A, B, \delta; x, y)$ over the probabilistic space $\mathcal{P} := (\Omega, \mathcal{A}, P)$, where:

- A is a given nonempty set and $B = \mathbb{C}$ or $B = \mathbb{R}$;
- $x\colon \Omega \to A$ and $y\colon \Omega \to B$ are functions defined on a given nonempty set Ω; they can be interpreted as experimental data of the regression model. Therefore we call them empirical data functions;
- $\delta\colon (\Omega \to B) \times (\Omega \to B) \to \overline{\mathbb{R}}$, is a function which can be interpreted as a deviation criterion of the theoretic functions from the empirical data. The δ function is defined through the probabilistic space $\mathcal{P}$ according to the classical concept of the least squares method.

For a given probabilistic regression structure $\mathfrak{P}$ one considers a family of all functions $\mathcal{F}$, called *theoretical functional model* for probabilistic regression structure $\mathfrak{P}$, which is a subset of the family of all functions acting from A to B. It is assumed that the class $\mathcal{F}$ is a linear functional model in the function space $(A \to B)$ with respect to the standard operations of adding and multiplying functions by a constant.

A natural question for a given probabilistic regression structure $\mathfrak{P}$ is the study and evaluation the optimal functions of the theoretic functional model $\mathcal{F}$, which are, with respect to the criterion δ, the best fitted to the empirical data, represented by the empirical data functions x and y. To be more specific, it is considered the extremal problem of determining and studying the class $\mathrm{Reg}(\mathcal{F}, \mathfrak{P})$ of all functions $f_0 \in \mathcal{F}$ minimizing the functional $\mathcal{F} \ni f \to \delta(f \circ x, y) \in \overline{\mathbb{R}}$. Each function $f_0 \in \mathrm{Reg}(\mathcal{F}, \mathfrak{P})$ is said to be the *regression function* in $\mathcal{F}$ with respect to $\mathfrak{P}$. The problem

of describing all regression functions in $\mathcal{F}$ with respect to $\mathfrak{P}$, is called the *regression problem* in $\mathcal{F}$ for $\mathfrak{P}$.

The first chapter contains a historical outline of the regression theory and a concise description asynchronous (ARS) and synchronous (SRS) regression structures. These concepts were borrowed from the article *Generalized approach to the problem of regression* (DOI 10.1007/s13324-014-0096-7) being an attempt of unification the theory of regression. It was published in 2015 and became motivation to write this doctoral dissertation. The second chapter presents the auxiliary facts about the integration of complex functions, used in the further part of the dissertation. In particular theorems on integration by substitution and on approximation of complex integrable functions by simple functions were proved.

The third chapter is an essential part of the dissertation. It provides the definition of the probabilistic regression structure $\mathfrak{P} := (A, B, \delta; x, y)$ over the probabilistic space $\mathcal{P} := (\Omega, \mathcal{A}, P)$. The main result here is Theorem 3.8 characterizing the class of regression functions. It results in a number of basic properties of regression functions, described in the later part. In particular, the conditions ensuring the uniqueness of the regression function were given. The first type of regression was described in terms of probabilistic regression structures. The relationship between real and complex probabilistic regression structures was also discussed. In particular, one presents an example justifying the consideration of complex probabilistic regression structures. Chapter four presents the problem of determining the regression functions in terms of a given finite basis of a linear set $\mathcal{F}$. The structure of the regression class and the procedure of determining the regression functions were discussed here. In addition, the regression problems for one-dimensional, two-dimensional and three-dimensional theoretical functional models $\mathcal{F}$ were solved. The general multidimensional case was treated briefly. The above considerations were illustrated with examples. One also discussed the transformation problem of regression function by replacing the sampling function y by the composition $g \circ y$, where g is a first order polynomial.

The fifth chapter presents a practical illustration of theoretical considerations. Based on Theorem 4.10, a numerical method of calculating the regression function for any given finite set of base functions was elaborated. There were also presented a real-life examples of determining the regression function based on the implementation of this method in MS Excel spreadsheet using the VBA programming language. The program code was attached to the appendix. Additionally, there were inserted screenshots of the numerical calculations for the real-life examples from chapter five.