Research on Teaching Method of Data Analysis Based on Statistical Regression

Peicheng Shi

Zhengzhou University of Industrial Technology Basic Education Department, Zhengzhou, Henan, 451100, China

spc_77@126.com

Keywords: Statistical Regression, Data Analysis, Teaching Methods.

Abstract: The background of the era of big data requires the undergraduates majoring in automation to have preliminary data analysis ability. On this basis, a data analysis case based on statistical regression is given, and the regression methods of Excel and MATLAB are analyzed in detail. At the same time, the regression method is given layer by layer when the data is returned. Students' computer practice proves that this teaching method can enable students to master the basic ideas and methods of data analysis and lay a good foundation for future research.

1. Introduction

"System modeling and simulation" is one of the compulsory courses for undergraduates majoring in modeling and simulation theory and technology in the school of automation, Beijing University of Aeronautics and Astronautics. System modeling and simulation technology is an important means for human to understand and improve the objective world[1]. It has important theoretical significance and practical value in many application fields such as aviation, aerospace, ship, automobile, etc. The teaching purpose of this course is to enable students to systematically understand the methods, technologies and application fields of system modeling and simulation, and train students to master the modeling methods and simulation tools of system simulation, as well as the design and implementation methods of simulation system. Through the study of the course, students are trained to understand and analyze the objective world objects from the perspective of methodology, epistemology and practice, have the ability to build object model system, design and construct simulation engineering system, and participate in the research and development of relevant engineering technology research and application system. Lay the necessary foundation. "System modeling and simulation" is an important course to cultivate students' basic skills in solving complex engineering problems. It is also one of the important skills for students to learn relevant professional courses and carry out scientific research in master's and doctor's stages. Knowledge driven modeling (such as Newtonian mechanics) and data-driven modeling are two main modeling methods[2]. The former includes classical analytical models in physics and mathematics, which can be modeled by ordinary differential, partial differential and matrix theory. In the past few years, with the development of artificial intelligence technology, people gradually pay attention to it. Truly rich data is the key to data-driven. Generally, neural networks and artificial intelligence need big data. It refers to a collection of data that can not be captured, managed and processed by traditional software tools in a certain time range. It needs a new processing mode. In order to have stronger decision-making power, insight and process optimization ability, massive, high growth rate and diversified information assets. Big data has the characteristics of large quantity and fast generation, which usually needs professional computer, mathematics and other theories and tools for processing and mining. In view of the importance of data-driven modeling and the enthusiasm of students to learn data model, and the lack of comprehensive data analysis training cases in the course of "system modeling and simulation", a data regression experiment on the passing rate of crowd evacuation simulation is designed[3]. The reason why crowd simulation is chosen as the main object of data analysis is that, on the one hand, because crowd simulation belongs to social simulation, students are familiar with its mode characteristics, but relatively new

DOI: 10.25236/icatpe.2020.188

academic research and other social forces and neural network modeling; on the other hand, through multiple volunteer group experiments, a large number of crowd trajectory data can be accumulated as the object of data analysis. Therefore, we choose crowd simulation experiment to teach students to use Excel, MATLAB and other tools for preliminary data regression analysis, which lays a good foundation for in-depth research in graduate stage.

2. Case Background: Statistical Regression Model of Gate Passing Rate

Generally, according to people's understanding of the model, the model can be divided into white box model, black box model and gray box model[4]. At present, human behavior is more of a black box model, such as the game in the stock market, the crowd behavior in the escape and evacuation, it is difficult to model with simple mathematical formula. But for this type of black box system, we can get the corresponding data, even big data in some way; then we can find the most suitable model through the statistical analysis of the data, and realize the black data regression modeling of box model. In fact, regression model is the most common model type of statistical analysis method. In the past 20 years, there are two main pedestrian behavior modeling methods: social force model and cellular automata model. The former is continuous while the latter is discrete. Generally, more calculation is needed to simulate social force model. With the improvement of computer computing, social power model has gradually become the mainstream because of its continuity. The biggest characteristic of social force model is its universality. It can be easily applied to any scene, but its micro behavior and macro statistics are far from the actual situation[5]. At present, more and more researches begin to focus on using neural network data-driven model to simulate crowd movement. Some scholars have proposed an artificial intelligence based human motion simulation method. This method needs to collect a large number of micro pedestrian movement data through road monitoring, and learn through neural network. The results show that the neural network can simulate the micro behavior of pedestrians when they cross the road. At the same time, some researches use recurrent neural network to complete the prediction of pedestrian trajectory in a specific area. Their research shows that the neural network model can simulate the behavior of pedestrians more accurately than the social force model. Different from other statistical methods, this course does not involve the mathematical principles and methods of regression analysis. It mainly uses a typical case, the evacuation exit capacity (EC), to teach students how to use statistical regression tools to solve practical problems[6]. Figure 1 is a screenshot of the simulation program for EC calculation using the developed crowd evacuation simulation software. The results in the digital simulation mode are shown in Table 1. Where, the unit of EC is the number of people per second, the unit of door width (W) is meter, and the unit of estimated pedestrian speed (V 0) is meter / second.

Table 1 Regression coefficient

Index variable	Non standardized coefficient		Standard coefficient	t	sig
	B	Standard error	β		
Constant	14630.934	4307.600		3.397	0.001
Gold reserve	13.084	2.807	0.322	4.660	0.000
Total Population	−0.296	0.058	−0.214	−5.063	0.000

3. Teaching Design of Statistical Regression Method

The first step is to teach students to visualize the data in Excel, mainly using Excel's insert scatter chart tool menu[7]. Because it is difficult to see the change trend of data simply by naked eyes, but it is more convenient to observe the change trend of data when visualizing to the curve chart, so as to realize the preliminary prediction of curve. That w is a constant value and EC changes with V. Each curve increases with the rise of V 0, but the trend of growth slows down, that is, it presents a reciprocal or logarithmic situation with small fluctuations[8]. Vis the constant value and EC changes with W. Each curve presents a good linear relationship, that is, it presents the situation of primary function with slight concave convex. Teach students to use MATLAB tools for fitting. With these basic skills of data processing, we can further teach students the method of neural network modeling. For example, the establishment of multi scene oriented artificial neural network (ANN) crowd movement model[9]. Here, the neural network model is still driven by data. It encapsulates the multi scene crowd behavior into a four layer neural network to output the speed and location of pedestrians. The training data and verification data used are all from real experiments. Comparing the simulation results with the real data, whether from the micro phenomenon or the macro data, if the designed neural network model is better than the social force model, then it can show the advantages of data modeling. At the same time, the input parameters of ANN crowd model can be added to the path planning information, and the input parameters can be rotated vector, which can be applied to various scenes, making up for the shortcomings of current neural network crowd motion model.

4. Conclusion

It can be seen from the above examples that the establishment of data analysis model is based on a large number of known actual data. First, analyze its basic characteristics from common sense and experience, analyze its most likely regression variables, and determine which regression variables and their forms (such as linear, quadratic, power index, logical Stryker curve, etc.) to take with the help of drawing (such as scatter diagram))And then it can be analyzed with data regression tools (such as least square method, lingo, matlab toolbox, etc.). If it is solved by software, it is necessary to conduct statistical analysis after data fitting, including R, F, bias, variance, etc., to evaluate the regression data model as a whole, so as to test whether the corresponding regression variables have significant impact on the dependent variables (if zero is included, it is not significant)[10]. If you are not satisfied with the result, you can continue to improve, such as changing a curve, adding a primary term, a secondary term and an interaction term. Through the observation and analysis of several senior students in the actual experimental work, in the actual process of using the computer, the differential requirements can be implemented according to the degree of students mastering the data regression skills. For students with certain foundation, only put forward the above functional requirements, point out the names of Excel and MATLAB related tools, and the rest of the regression can be required to be realized by students themselves; when students encounter programming difficulties, teachers can give certain prompts. For students with poor foundation, it is better to give specific analysis steps, and the tool software must also prompt to the corresponding menu; at the same time, whenever they encounter difficulties, take the way of discussion and further explanation until the students understand and realize the data regression function. It has been proved that the data regression analysis cases designed in this paper can be effectively applied to large-scale assignments or experimental cases of undergraduates majoring in modeling and simulation. It is an effective teaching method to help students understand the concept of data analysis and master the corresponding data analysis skills in this course. After several rounds of application practice, the teaching team of Beihang automation college embodies the effectiveness of this method in classroom interaction, student feedback, course evaluation and subsequent graduation design practice. The above cases will be further improved in the future, so that they can be more effectively applied to the excellent courses being constructed by the school, and provide tools and platform support for further improving students' practical ability and scientific research innovation ability.

References

[1] Kruppa, J. STATISTICAL REGRESSION AND CLASSIFICATION Norman Matloff. Boca Raton, FL: Chapman & Hall/ CRC Press. 528 pages, 2018.

[2] Tsung-Shan, Tsou. A robust likelihood approach to inference about the kappa coefficient for

correlated binary data. Statistical Methods in Medical Research, vol. 28, no. 4, 2018.

[3] Stöckl, Dietmar., Dewitte, Katy., Thienpont, Linda, M. Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data?. Clinical Chemistry, no. 11, pp. 11, 2020.

[4] Qoua, L. Her., Jessica, M. Malenfant., Sarah, Malek. A Query Workflow Design to Perform Automatable Distributed Regression Analysis in Large Distributed Data Networks. Egems, vol. 6, no. 1, 2018.

[5] A Islamiyati, Fatmawati, N Chamidah. Estimation of Covariance Matrix on Bi-Response Longitudinal Data Analysis with Penalized Spline Regression. Journal of Physics Conference Series, vol. 979, no. 1, pp. 012093, 2018.

[6] Jr Eduardo E. Ribeiro, Walmes M. Zeviani, Wagner H. Bonat, Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Data. Statistical Modelling, 2018.

[7] Belkais Altendji, Jacques Demongeot, Ali Laksaci, Functional data analysis: Estimation of the relative error in functional regression under random left-truncation. Journal of Nonparametric Statistics, vol. 30, no. 2, pp. 472-490, 2018.

[8] Rodbard, David. Statistical Quality Control and Routine Data Processing for Radioimmunoassays and Immunoradiometric Assays. Clinical Chemistry, no. 10, pp. 10, 2020.

[9] Linnet, Kristian. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. Clinical Chemistry, no. 5, pp. 5, 2020.

[10] Kym, Ie. Snell., Joie, Enso., Thomas, Pa. Debray. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures?. Statistical Methods in Medical Research, vol. 27, no. 11, 2017.