DOCTORAL THESIS

Checking the adequacy of regression models with complex data structure

Guo, Xu

Date of Award:
2014

Abstract

In this thesis, we investigate the model checking problem for parametric regression model with missing response at random and nonignorable missing response. Besides, we also propose a hypothesis-adaptive procedure which is based on the dimension reduction theory. Finally, to extend our methods to missing response situation, we consider the dimension reduction problem with missing response at random.

The first part of the thesis introduces the model checking for parametric models with response missing at random which is a more general missing mechanism than missing completely at random. Different from existing approaches, two tests have normal distributions as the limiting null distributions no matter whether the inverse probability weight is estimated parametrically or nonparametrically. Thus, p-values can be easily determined. This observation shows that slow convergence rate of nonparametric estimation does not have significant effect on the asymptotic behaviours of the tests although it may have impact in finite sample scenarios. The tests can detect the alternatives distinct from the null hypothesis at a nonparametric rate which is an optimal rate for locally smoothing-based methods in this area. Simulation study is carried out to examine the performance of the tests. The tests are also applied to analyze a data set on monozygotic twins for illustration.

In the second part of the thesis, we consider model checking for general linear regression model with non-ignorable missing response. Based on an exponential tilting model, we first propose three estimators for the unknown parameter in the general linear regression model. Three empirical process-based tests are constructed. We discuss the asymptotic properties of the proposed tests under null and local alternative hypothesis with different scenarios. We find that these three tests perform the same in the asymptotic sense. Simulation studies are also carried out to assess the performance of our proposed test procedures.

In the third part, we revisit traditional local smoothing model checking procedures. Noticing that the general nonparametric regression model can be considered as a special multi-index model, we propose an adaptive testing procedure based on

the dimension reduction theory. To our surprise, our method can detect local alternative at faster rate than the traditional optimal rate. The theory indicates that in model checking problem, dimensionality may not have strong impact. Simulations are carried out to examine the performance of our methodology. A real data analysis is conducted for illustration.

In the last part, we study the dimension reduction problem with missing response at random. Based on the work in this part, we can extend the adaptive testing procedure introduced in the third part to the missing response situation. When there are many predictors, how to efficiently impute responses missing at random is an important problem to deal with for regression analysis because this missing mechanism, unlike missing completely at random, is highly related to high-dimensional predictor vector. In sufficient dimension reduction framework, the fusion-refinement (FR) method in the literature is a promising approach. To make estimation more accurate and efficient, two methods are suggested in this paper. Among them, one method uses the observed data to help on missing data generation, and the other one is an ad hoc approach that mainly reduces the dimension in the nonparametric smoothing in data generation. A data-adaptive synthesization of these two methods is also developed. Simulations are conducted to examine their performance and a HIV clinical trial dataset is analysed for illustration.

Keywords: Model checking; Inverse probability weight; Non-ignorable missing response; Adaptive; Central subspace; Dimension reduction; Data-adaptive Synthesization; Missing recovery; Missing response at random; Multiple imputation.

Acknowledgements

First and foremost, I would like to take this opportunity to express my great appreciation to my supervisor Prof. ZHU Lixing. Without his support and guidance, this work would not have been completed. Learning many different things from him in the fields of statistics, mathematics and research, and experiencing his diverse excellent qualities during the last years are what I will be ever thankful to him. I would also like to thank my co-supervisor Dr. TONG Tiejun, for his invaluable advice and helpful comments.

Also I would like to thank Prof. WONG Wing-Keung in the economics department of Hong Kong Baptist University. Without his continuous guidance and encouragement, I will not enter the economic area and enjoy much theoretical and empirical knowledge in economics and finance.

I wish to thank other faculty members in the mathematics department, especially CHUI Claudia, YUM Rainbow, LAM Tammy, YEUNG C. W., HUI Vicky and LI Candy for their excellent help. Also I would like to thank LO Kamfai of Graduate School for his great help.

Many people have helped and guided me–both professionally and personally–in the last few years. Prof. LIN Lu, Dr. XU Wangli, LI Qiuyue, YANG Yiping, PENG Heng and LI Gaorong have always been supportive and instructive. I am very grateful to Dr ZHU Liping, WU Jianhong, LI Zaixing, WU Ping, FANG Yun, YU Zhou, FENG Zhenghui, ZHANG Jun, FAN Yan, WANG Tao, XU Peirong, WANG Cheng, XIA Qiang, ZHOU Jingke, ZHU Xuehu and all the people of the Lixing research team for various combinations of help, support and inspiration. Special thanks go to TIAN Qiushi for her tremendous assistance and encouragement throughout my graduate career.

Finally, I wish to express my gratitude to my wife, NIU Cuizhen. I'm very happy and lucky to be married to this wonderful woman. Without her great support and

love, I would not complete this work smoothly. I also wish to thank my parents for their constant support and encouragement, both emotionally and intellectually.

Table of Contents