

DOCTORAL THESIS

Checking the adequacy of regression models with complex data structure

Guo, Xu

Date of Award:
2014

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

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

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	vi
List of Tables	ix
List of Figures	xi
Chapter 1 Introduction	1
1.1 Model Checking for Regression	2
1.1.1 Smoothing-Based Tests	2
1.1.2 Tests Based on Empirical Regression Processes	5
1.2 Sufficient Dimension Reduction in Regression	6
1.2.1 Sliced Inverse Regression	7
1.2.2 Minimum Average Variance Estimation	8
1.3 Outline of the Thesis	10
Chapter 2 Model Checking for Parametric Regressions with Response	
Missing at Random	12
2.1 Introduction	12

2.2	Test Procedures	15
2.2.1	Construction of Test Statistics	15
2.2.2	Asymptotic Behavior of the Test Statistics	18
2.3	Numerical Analysis	22
2.3.1	Simulation Study	22
2.3.2	Real Data Analysis	29
2.4	Discussion	29
2.5	Appendix. Proofs of Theorems	31

Chapter 3 Model Checking for General Linear Regression with Nonignorable Missing Response **47**

3.1	Introduction	47
3.2	Construction of Test Statistics	50
3.3	Asymptotic Behavior of the Test Statistics	53
3.3.1	Asymptotic Properties with Known γ^*	53
3.3.2	Asymptotic Properties with Estimated $\hat{\gamma}$ from Independent Survey	54
3.3.3	Asymptotic Properties with Estimated $\hat{\gamma}$ from Validation Sample	55
3.3.4	Monte Carlo Approximation	56
3.4	Simulation Study	59
3.5	Appendix. Proofs of Theorems	61

Chapter 4 Model Checking for Generalized Linear Models: An Hypothesis-Adaptive Method **79**

4.1	Introduction	79
4.2	Adaptive Test Procedure	81
4.2.1	Review of DEE	83
4.2.2	Review of MAVE	84
4.2.3	Estimation of Dimension d	85

4.3	Asymptotic Properties	86
4.3.1	Power Study	87
4.4	Numerical Analysis	88
4.4.1	Simulations	88
4.4.2	Real Data Analysis	94
4.5	Discussion	94
4.6	Appendix. Proof of the Theorems	97
Chapter 5 Dimension Reduction with Missing Response at Random		117
5.1	Introduction	117
5.2	Semiparametric Dimension Reduction Assisted Recovery	121
5.2.1	Selection Probability Assisted Recovery	122
5.2.2	Complete Case Assisted Recovery	123
5.3	SIR with Missing Response	124
5.3.1	Application of SIR to SPAR and CCAR	124
5.3.2	Determination of the Structural Dimension	126
5.4	Simulation Studies	126
5.4.1	Estimation of the Central Subspace	128
5.4.2	Data-Adaptive Synthesization	129
5.5	Application to A HIV Dataset	130
5.6	Conclusion	132
5.7	Appendix. Proof of the Theorems	132
Bibliography		143
Curriculum Vitae		153