DOCTORAL THESIS

Multivariate statistical diagnostics with application to the growth curve model

Pan, Jianxin

Date of Award:
1996

Multivariate Statistical Diagnostics with Application to the Growth Curve Model

PAN Jian-Xin

A thesis submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

April 1996

Hong Kong Baptist University

PREFACE

The growth curve model (GCM) is a generalized multivariate analysis-of-variance model (GMANOVA), which was first proposed by Potthoff and Roy (1964) and has been studied subsequently by many authors such as Rao (1965), Khatri (1966) and von Rosen (1989). This model is useful especially for investigating growth problems on short times in economics, biology, medical research and epidemiology. Also, it is one of the fundamental tools for dealing with longitudinal data especially with serial correlation (Jones, 1993) as well as repeated measures (Crowder and Hand, 1990). However, it is not uncommon to find outliers and influential observations in growth data that significantly affect the estimates of the GCM. Under a specific covariance structure, namely, spherical covariance structure, Liski (1991) presented several measures for detecting outliers and influential observations in the GCM. For this model with other commonly used covariance structures, such as Rao's simple covariance structure (SCS) as well as unstructured covariance scheme (UCS), how multiple outliers and/or influential observations should be detected effectively is still open.

It is the purpose of this thesis to systematically investigate multivariate statistical diagnostics for the GCM. The research will be mainly concentrated on the GCM with the two covariance structures mentioned above. The multivariate diagnostic techniques used in this thesis can be classified into two categories: global influence (also known as case-deletion approach) and local influence, and each of them is employed to diagnose the adequacy of the GCM based on likelihood framework as well as Bayesian framework.

Chapter 1 of this thesis gives a background of statistical diagnostics, a brief introduction to multiple outlier identification in high-dimensional data sets, and a brief review of the GCM as well as its model selection criteria with respect to covariance structure, which serves as a basis of diagnostics development for the GCM. Three real-life biological data sets are analyzed in order to emphasize the use of the selection criteria. Also, the main contributions of this thesis on statistical diagnos-

tics to the GCM are presented in a summarized form in this chapter. In addition, some preparations related to matrix derivatives and matrix-variate distributions are provided for later use.

In Chapter 2, based on likelihood framework, we use case-deletion approach to explore the relation between the multiple individual deletion model and the mean shift regression model, to build up multiple outlier detection criteria, and to construct influential measures based on the generalized Cook's distance and the confidence ellipsoid volume. Also, the influential measures are used to assess a linear combination of the regression coefficients. These diagnostic techniques are applied to the GCM with SCS and UCS, respectively. For illustration, the three biological data sets used in Chapter 1 are analyzed by the use of such techniques for outlier detection and influential observation identification.

Chapter 3 is devoted to discussion of Cook's (1986) likelihood local influence applied to diagnose the adequacy of the GCM with SCS and UCS, respectively. Under these two covariance structures, the observed information matrix as well as the Hessian matrix for the GCM are investigated in detail, which serves as a basis of the local influence assessment for the model. As an ancillary result, the Hessian matrix is shown to be invariant under a one-to-one measurable transformation of parameters. Also, the real-life data sets are analyzed by using the local influence approach.

Based on Bayesian framework, in Chapter 4 we discuss the influence of a subset of observations on the growth fittings by the use of case-deletion approach. Under a non-informative prior distribution, the posterior distributions of the parameters for the GCM with SCS and UCS are derived, respectively. The Kullback-Leibler divergence or Bayesian entropy in information theory is then used for measuring the change of the posterior distributions when the subset of observations is deleted from the data. Also, the numerical examples presented in the pervious chapters are analyzed again by using the methods of this chapter.

From the Bayesian point of view, Chapter 5 is devoted to discussion of local influence approach for the GCM. The fundamental idea is to replace the likelihood displacement with the Kullback-Leibler divergence. For the two commonly used covariance structures, SCS and UCS, the Bayesian Hessian matrices of the GCM are investigated respectively under an abstract perturbation scheme, which play pivotal roles in the Bayesian local influence for the model. Also, some new properties of the Bayesian Hessian matrix are obtained as ancillary results. For illustration, the covariance-weighted perturbation is considered especially and employed to analyze the real-life biological data sets addressed previously.

In addition, more comments and recommendations for further research on this subject are proposed, which could be viewed as a guideline of further investigation for the GCM. Finally, the subroutines for calculating the model selection criteria with respect to covariance structure and all the diagnostic measures obtained in this thesis, written in MATLAB, are presented in Appendix of this dissertation.

Jian-Xin Pan
Department of Mathematics
Hong Kong Baptist University
224 Waterloo Road, Kowloon Tong
Hong Kong

April 15, 1996

CONTENTS