Mathematics Monograph Series 8

Growth Curve Models and Statistical Diagnostics

Jianxin Pan and Kaitai Fang

（生长曲线模型及其统计诊断）

SCIENCE PRESS
Beijing

Mathematics Monograph Series 8

Growth Curve Models and Statistical Diagnostics

Jianxin Pan and Kaitai Fang

Mathematics Monograph Series 8

Jianxin Pan and Kaitai Fang

Growth Curve Models and Statistical Diagnostics

(生长曲线模型及其统计诊断)

Responsible Editor: Jiashan Liu, Deping Yan, Yang Zhang

Published by Science Press
16 Donghuangchenggen North Street
Beijing 100717, China

Printed in Beijing

ISBN 978-7-03-019532-6 (Beijing)

To the memory of my mother, to my father in his 70^{th} year, and to my wife Haiyan and my son Kainan for their patience during the writing of this book.

J.X. Pan

Jianxin Pan
Centre for Medical Statistics
Department of Mathematics
Keele University
Staffordshire, ST5 5BG
United Kingdom

To my wife Tingmei and my two daughters, Ying and Yan, for their constant support.

K.T. Fang

Kaitai Fang
Department of Mathematics
Hong Kong Baptist University
224 Waterloo Road, Kowloon
Hong Kong

Preface

A growth curve model (GCM) is a generalized multivariate analysis-of-variance model (GMANOVA), first summarized by Potthoff and Roy (1964) and studied subsequently by many authors including Rao (1965), Khatri (1966) and von Rosen (1989). The GCM is especially useful for investigating grwoth problems on short time series in areas such as economics, biology, medical research, and epidemiology. It is also a fundamental tool for analyzing longitudinal data especially with serial correlation (Jones, 1993) and repeated measures (Crowder and Hand, 1990). It is not uncommon, however, to find outliers and influential observations in growth data that significantly affect statistical inference in the GCM.

The purpose of this book to introduce the theory of the GCM with particular emphasis on statistical diagnostics, which is mainly based on recent work on diagnostics made by the authors and their collaborators. This book is intended for researchers who are working in the area of theoretical studies related to the GCM as well as multivariate statistical diagnostics, and for applied statisticians working in application of the GCM to practical areas. Hence, on the one hand, we provide theoretical proofs for the most theorems in this book. On the other hand, applications of these techniques to practical data analysis are emphasized; for example, almost every approach discussed in this book is illustrated with practical examples. In addition, the computer programmes for calculating various measurements involved in this book have been written in S-PLUS and GENSTAT. We will put the computer programmes on our web site in due course. A link to the web site can be found in the list of author web pages at the Springer web page, *www.springer-ny.com*.

The statistical diagnostics considered in this book focuses mainly on GCMs with two specific covariance structures, namely, *Rao's simple covariance structure* (SCS) and *unstructured covariance* (UC), since these two covariance structures are very common in practice and some other covariance structures are their special cases. For example, the *uniform covariance* *structure* and *random-effects covariance structure* are two special cases of SCS. GCMs with other covariance structures can also be analyzed in a similar manner. The multivariate diagnostic techniques addressed in this book are classified into two categories: *global influence* (also known as *case-deletion* approach) and *local influence*; and each of these is used to diagnose the adequacy of GCMs within the likelihood and Bayesian frameworks as well.

Chapter 1 of this book gives a background of statistical diagnostics, a brief introduction to multiple outlier identification in multivariate data sets, a brief review of the GCM, and the related model selection criteria with respect to covariance structure. Also, the main approaches and results on statistical inferences and diagnostics in GCMs are presented in a summarized form in this chapter. In addition, preparatory materials related to matrix derivatives and matrix-variate distributions are provided for later use.

In Chapter 2, the fundamental concepts of GCMs are introduced and several most commonly encountered forms of the models are explained in terms of practical examples in biology, agriculture and medical sciences. The generalized least square estimate (GLSE) and admissibility of estimates on linear combinations of regression coefficients are discussed. We show that the GLSE of the regression coefficient is also the best linear unbiased estimate (BLUE) in the sense of matrix loss functions. We also study here the necessary and sufficient conditions of admissible estimates of linear combinations of the regression coefficients.

Maximum likelihood estimate (MLEs) of the regression coefficient and dispersion component in growth curve models are discussed in Chapter 3. We also study the expectation and variance-covariance matrix of the estimates. In general, the MLE of the regression coefficient is different from the GLSE given in Chapter 2. In fact, the latter is a linear function of the response variable while the former is not. There is indeed a special case, however, in which the MLE is completely identical to the GLSE. In this case, statistical inferences based on the MLE in growth curve models becomes simpler. This special case is nothing but SCS, in which the dispersion component matrix Σ is decomposed as two orthogonal components. This point will be shown with illustrative examples. As an alternative to the MLE, restricted maximum likelihood (REML) estimates are studied in the context of growth curve models with SCS and random effects covariance structure in this chapter. Estimates of the dispersion components are unbiased in this case. Numerical studies are conducted to compare the GLSE, MLE and REML in growth curve models.

Within the likelihood framework in Chapter 4 we use the case deletion technique to explore the relationship between the multiple individual deletion model (MIDM) and the mean shift regression model (MSRM), to build up multiple outlier detection criteria, and to construct influence measurements based on the generalized Cook's distance and the confidence ellipsoid volume. Also, influence measurements are used to assess a linear combination of regression coefficients. These diagnostic techniques are applied to GCMs with SCS and UC, respectively. For illustration, some biological, medical, and agricultural data sets are analyzed using these diagnostic techniques for outlier detection and influential observation identification.

Chapter 5 is devoted to discussing how Cook's (1986) likelihood-based local influence technique could be used to diagnose the applied of GCMs with SCS and UC, respectively. With these two specific covariance structures, the observed information matrix and the Hessian matrix are studied; the Hessian matrix serves as a basis of the local influence assessment in these models. As an ancillary result, the Hessian matrix is shown to be invariant under a one-to-one measurable transformation of parameters. Also, the practical data sets analyzed in the previous chapters are reanalyzed using local influence approach discussed in this chapter.

In Chapter 6, within the Bayesian framework, we discuss the influence of a subset of observations on growth fittings in terms of case deletion technique. Under a noninformative prior distribution, the posterior distributions of the parameters in GCMs with SCS and UC are considered, respectively. The Kullback-Leibler divergence is used to measure the change of posterior distributions when the subset of observations is removed from the model. The numerical examples addressed in the previous chapters are analyzed once again using the methods developed in this chapter.

Chapter 7 is devoted to discussion of the local influence approach in the GCM from a Bayesian point of view. The fundamental idea of *Bayesian local influence* is to replace the likelihood displacement of Cook's local influence with the Kullback-Leibler divergence. For the two commonly used covariance structures, SCS and UC, Bayesian Hessian matrices in the GCM are studied under an abstract perturbation. Those matrices play a pivotal role in the Bayesian local influence. Also, some properties of Bayesian Hessian matrix are considered as ancillary results, and the relationships between likelihood-based local influence and Bayesian local influence are studied. For illustration, the covariance-weighted perturbation is considered especially and employed to analyze several practical data sets.

We would like to thank the following people from Hong Kong Baptist University: Prof. C. F. Ng, Dr. C. W. Tse, and Prof. F. J. Hickernell. We also thank Prof. G. MacKenzie and Prof. P. W. Jones of Keele University, Prof. R. Thompson of IACR-Rothamsted, and Prof. X. R. Wang of Yunnan University. All gave valuable inspiration on our research and constant encouragement. We are grateful to Prof. P. M. Bentler and Prof. D. von Rosen for their encouragement in the writing of this book. We also acknowledge with gratitude P.M. Bentler, D. von Rosen, R. Thompson, G. MacKenzie, and anonymous referees for their reading of the manuscript and their invaluable comments and suggestions. We thank our various coauthors in a series of papers, including D. von Rosen, E.P. Liski, W.K. Fung, and P. Bai for the nice and very stimulating collaborations, from which some results are reflected in this book. Also, we are indebted to J. S. Liu and X. Y. Ge of Science Press and J. Kimmel and J. Wolkowicki of Springer-Verlag for their hard work on the publication of this book. Mrs J.Drewery of Keele University kindly helped us to prepare the Index pages and her great help is very appreciated. Last but not least, we gratefully acknowledge support from our families.

Jian-Xin Pan's research was partially supported by a grant from Yunnan Science Foundation, a fellowship from Hong Kong Baptist University during his time there as a Ph. D. student, a grant from Agriculture, Environment and Fisheries Department of the Scottish Office during his time as a post-doctoral research fellow in IACR-Rothamsted, and a grant from the Acute (NHC) Trust of North Staffordshire. Kai-Tai Fang's research was partially supported by the Hong Kong University Grant Council. We would like to acknowledge gratefully the generous support of all these institutions.

Keele University — Jian-Xin Pan

Hong Kong Baptist University — Kai-Tai Fang

21st February, 2007

Acronyms

AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
BLUE	Best Linear Unbiased Estimate
BLUP	Best Linear Unbiased Prediction
CLS	Cyclic Lattice Squares
DFFITS	Difference between fitted values
EM	Expectation–Maximization algorithm
GCM	Growth Curve Model
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GLP	Good Lattice Points
GLSE	Generalized Least Square Estimate
GMANOVA	Generalized Multivariate Analysis-of-Variance
i.i.d.	Independent Identically Distributed random variable
KLD	Kullback–Leibler Divergence
LISREL	Linear structure regression equation analysis
LSE	Least Square Estimate
MAD	Median of Absolute Deviation from the median
MCMC	Markov Chain Monte Carlo
MD	Mahalanobis Distance
MIDM	Multiple Individual Deletion Model
MLE	Maximum Likelihood Estimate
MSRM	Mean Shift Regression Model
MVE	Minimum Volume Ellipsoid estimate
OMRM	Ordinary Multivariate Regression Model
p.d.	Positive Definite matrix
p.d.f.	Probability Density Function
PP	Projection Pursuit
PQL	Penalized Quasi-Likelihood estimator
REML	Restricted Maximum Likelihood
RSS	Residual Sum of Squares
SCS	Rao's Simple Covariance Structure
SVD	Singular Value Decomposition
UC	Unstructured Covariance

Notations

Matrices are denoted by upper case bold letters, column vectors by lower case bold letters, and scalars by lower case letters. For example, $\boldsymbol{X}$, $\boldsymbol{x}$ and x represent a matrix, a column vector, and a scalar, respectively.

Space

R^p	p-dimensional Euclidian space
S^q	unit sphere in the q-dimensional Euclidian space R^q

Scalar

$J(\boldsymbol{X} \to \boldsymbol{Y})$	Jacobian of the transformation from $\boldsymbol{X}$ to $\boldsymbol{Y}$
$\mathrm{tr}(\boldsymbol{A})$	trace of the square matrix $\boldsymbol{A}$
$\det(\boldsymbol{A})$	determinant of the square matrix $\boldsymbol{A}$
$\Gamma_m(a)$	multivariate gamma function, i.e., $\Gamma_m(a) = \pi^{m(m-1)/4} \prod_{i=1}^m \Gamma(a-(i-1)/2)$

Vector

$\mathbf{1}_p$	$\mathbf{1}_p = (1, 1, \cdots, 1) \in R^p$, i.e., the p-variate vector of ones
$\boldsymbol{d}_{\max}$	unit eigenvector associated with the largest absolute eigenvalue of the Hessian matrix
$\mathrm{vec}(\boldsymbol{A})$	direct operator of the matrix $\boldsymbol{A}$
$\mathrm{vec}'(\boldsymbol{A})$	transpose of the vector $\mathrm{vec}(\boldsymbol{A})$
$\mathrm{svec}(\boldsymbol{A})$	symmetric direct operator of the symmetric matrix $\boldsymbol{A}$
$\mathrm{svec}'(\boldsymbol{A})$	transpose of the vector $\mathrm{svec}(\boldsymbol{A})$
$\mathrm{diag}(\boldsymbol{A})$	vector formed by the diagonal elements of $\boldsymbol{A}$

Matrix

$\boldsymbol{A}'$	transpose of the matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	inverse of the nonsingular matrix $\boldsymbol{A}$
$\boldsymbol{A}^{+}$	Moore–Penrose generalized inverse of the matrix $\boldsymbol{A}$
$\boldsymbol{I}_p$	identity matrix with order $p \times p$

$\boldsymbol{E}_{ij}(p,q)$	$p \times q$ matrix with (i,j)th element being one and others being zero. in short, denoted $\boldsymbol{E}_{ij}$
$\boldsymbol{K}_{pq}$	permutation matrix with order $pq \times pq$; for $p = q$, denoted $\boldsymbol{K}_{p^2}$
$\boldsymbol{S}_p$	duplication matrix with order $p^2 \times p(p+1)/2$
$\boldsymbol{P_A}$	$\boldsymbol{P_A} = \boldsymbol{A}(\boldsymbol{A}^\tau\boldsymbol{A})^{-1}\boldsymbol{A}^\tau$, the projection matrix of $\boldsymbol{A}$
$\boldsymbol{Q_S}$	$\boldsymbol{Q_S} = \boldsymbol{SQ}(\boldsymbol{Q}^\tau\boldsymbol{SQ})^{-1}\boldsymbol{Q}^\tau$, a semiprojection matrix.
$\boldsymbol{A} \otimes \boldsymbol{B}$	Kronecker product of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$
$\boldsymbol{A} * \boldsymbol{B}$	Hadamard product of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$
$\boldsymbol{\Sigma} > \boldsymbol{0}$	$\boldsymbol{\Sigma}$ is a positive definite matrix
$\boldsymbol{\Sigma}^{1/2}$	square root of the matrix $\boldsymbol{\Sigma}$
$\mathrm{Cov}(\boldsymbol{X})$	variance-covariance matrix of the random matrix $\boldsymbol{X}$, i.e., $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{Cov}(\mathrm{vec}(\boldsymbol{X}))$

Univariate distribution

$N(\mu, \sigma^2)$	univariate normal distribution with expectation μ and variance σ^2
t_p	Student's $t$distribution with p degrees of freedom
χ^2_p	chi-square distribution with p degrees of freedom
$F_{n,m}$	F distribution with n and m as first and second degrees of freedom, respectively
$\mathrm{Gamma}(a,b)$	gamma distribution with parameters a and b
$\mathrm{Beta}(a,b)$	beta distribution with parameters a and b
$\Lambda(p,m,n)$	Wilk's distribution with three parameters p, m, and n
$GT^2(m,r,n)$	Hotelling's generalized T^2 distribution with three parameters m, r, and n

Vector-variate distribution

$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	p-dimensional normal distribution with expectation vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma} > \boldsymbol{0}$
$t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$	p-dimensional t distribution with location $\boldsymbol{\mu}$ and dispersion $\boldsymbol{\Sigma} > \boldsymbol{0}$, and ν degrees of freedom

Matrix-variate distribution

$N_{p,n}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$	matrix-variate normal distribution with location matrix $\boldsymbol{M}$ and dispersion matrices $\boldsymbol{\Sigma} > \boldsymbol{0}$ and $\boldsymbol{\Omega} > \boldsymbol{0}$
$W_m(n, \boldsymbol{\Sigma})$	Wishart distribution with parameters n and $\boldsymbol{\Sigma} > \boldsymbol{0}$
$t_{p,n}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}, \nu)$	matrix-variate t distribution with location matrix $\boldsymbol{M}$ and dispersion matrices $\boldsymbol{\Sigma} > \boldsymbol{0}$ and $\boldsymbol{\Omega} > \boldsymbol{0}$ and ν degrees of freedom

Contents

Chapter 1

Introduction

Statistical diagnostics is one of the most useful techniques in statistical science. The aim of diagnostics is to detect outliers that deviate from the postulated model, to identify influential observations that have large effects on the statistical inference drawn from the postulated model, and to validate the chosen statistical model. The main theme of this book is to comprehensively explore multivariate diagnostic techniques, which are specifically suitable for diagnosing the adequacy of multivariate models, with particular emphasis on the application to growth curve models. The approaches employed are case-deletion and local influence within the likelihood and Bayesian frameworks. We give a brief introduction to statistical diagnostics in Section 1.1 and the associated multivariate techniques in Section 1.2. Section 1.3 is devoted to a brief review of growth curve models as well as model selection criteria with respect to covariance structures. In Section 1.4, the main approaches and results in this book on statistical diagnostics for growth curve models are outlined in a summarized form. Some preparatory materials related to matrix derivatives and matrix-variate distributions are given in Section 1.5 for later use.

1.1 General Remarks

1.1.1 Statistical diagnostics

In statistical science, statistical models play an important role in analyzing data, making statistical inferences, and making future predictions. As long as a random sample or data set is drawn from a practical phenomenon, statisticians often need to build up a "good" statistical model for fitting the