ENHANCING MULTIVARIATE LINEAR MIXED MODEL WITH TWO-LEVEL RANDOM EFFECTS FOR LONGITUDINAL DATA AND ITS APPLICATION TO PISA AND NATIONAL EXAMINATION SCORES

VERA MAYA SANTI

STATISTICS AND DATA SCIENCE
FACULTY OF MATHEMATICS AND NATURAL SCIENCE
IPB UNIVERSITY
BOGOR
2024

STATEMENT REGARDING DISSERTATION AND INFORMATION SOURCE AND COPYRIGHTS

I hereby affirm that the dissertation titled " Enhancing Multivariate Linear Mixed Model with Two-Level Random Effects for Longitudinal Data and Its Application to PISA and National Examination Scores" is entirely my own work, supported by my supervisory committee. This work has not been presented in any format to any other academic institution. All references to information derived from published or unpublished manuscripts are duly acknowledged within the dissertation and listed in the references section. Consequently, I assign the copyright of my dissertation to Bogor Agricultural University, Bogor, Indonesia.

Bogor, July 2024

Vera Maya Santi
NIM G161180021

RINGKASAN

VERA MAYA SANTI. Enhancing Multivariate Linear Mixed Model with Two-Level Random Effects for Longitudinal Data and Its Application to PISA and National Examination Scores. Dibimbing oleh KHAIRIL ANWAR NOTODIPUTRO, INDAHWATI, M.SI dan BAGUS SARTONO.

Perluasan cakupan untuk menambahkan peubah respon multivariat dalam analisis data *cross-sectional* menjadi penting ketika mempelajari fenomena yang saling berhubungan dan multidimensi. Penambahan peubah respon multivariat memungkinkan peneliti untuk mengeksplorasi hubungan dan ketergantungan antara beberapa peubah respon secara simultan. Kebutuhan untuk menganalisis multivariat respons dalam data *cross-sectional* muncul ketika peneliti bertujuan untuk memahami pola yang rumit, ketergantungan, dan interaksi di antara berbagai faktor dalam populasi tertentu.

Pada kondisi riil di lapangan, sekumpulan observasi seringkali diukur lebih dari satu kali, fenomena ini biasa disebut dengan pengukuran berulang. Pengukuran berulang, melibatkan penilaian suatu objek penelitian pada berbagai waktu pengamatan. Data hasil pengukuran berulang membentuk struktur data longitudinal. Beberapa peneliti telah mengembangkan pendekatan alternatif untuk menganalisis data longitudinal. Pendekatan yang lebih umum salah satunya adalah model campuran linier/*linear mixed model* (LMM).

Masalah lain dalam data longitudinal adalah adanya peubah respon multivariat dimana lebih dari satu respon diamati dalam survei. Salah satu cara untuk mengatasi hal ini adalah dengan memodelkannya secara serentak menggunakan model campuran linier multivariat. Kompleksitas kasus data riil akan meningkat ketika efek acak bersifat multilevel atau hierarki. Pertanyaan penelitian substantif yang muncul dalam struktur data longitudinal dimana variabel respon bersifat multivariat, dan efek acak bersifat multilevel, melibatkan isu-isu seperti perubahan yang saling terkait antar peubah respon.

Contoh data berhirarki dengan sejumlah amatan adalah skor data PISA dan juga data nilai ujian nasional (UN) Sekolah Menengah Atas (SMA). Penelitian ini dilakukan berulang kali untuk mengamati kecendrungan rata-rata skor PISA siswa Indonesia dan rata-rata nilai ujian nasional di level SMA. Data skor PISA dirancang untuk memasukkan sampel yang representatif dari populasi target yang merupakan hasil survei yang dilakukan oleh lembaga OECD. Terdapat tiga mata uji pada survei OECD yaitu literasi membaca, matematika dan sains. Sedangkan data rata-rata nilai ujian nasional SMA bersumber dari Data Pokok Pendidikan (DAPODIK).

Studi *cross-sectional* juga telah dilakukan untuk menganalisis skor data PISA. Dalam kasus ini, sangat penting untuk menangkap apakah ada asosiasi atau hubungan antara skor PISA siswa antar tiga mata uji dan juga sekolah sebagai pengaruh acaknya. Selanjutnya studi longitudinal juga telah dilakukan. Studi ini diterapkan pada data nilai rata-rata UN SMA jurusan IPA di provinsi Jawa Barat. Dalam studi longitudinal ini, sangat penting untuk menangkap tren antar nilai rata-rata UN untuk enam mata pelajaran yaitu bahasa Indonesia, Bahasa Inggris, matematika, fisika, kimia dan biologi dari waktu ke waktu. Hal penting lainnya yaitu untuk menangkap apakah ada kontribusi sekolah sebagai pengaruh acak terhadap peubah respon.

Analisis empiris *multivariate linear mixed model* (MLMM) telah dilakukan untuk struktur data *cross-sectional* maupun longitudinal. Meskipun skor data PISA menjadi subjek dari beberapa penelitian di Indonesia, tetapi analisis secara simultan dari ketiga skor data PISA tersebut menggunakan MLMM belum pernah dilakukan. Begitu pula analisis secara simultan terhadap rata-rata nilai UN SMA IPA di Jawa Barat dengan struktur data longitudinal serta penambahan pengaruh acak menggunakan MLMM. Hasilnya menunjukan bahwa dengan penambahan pengaruh acak di dalam model, mampu menurunkan standar eror dari penduga parameter tetap. Selain itu, untuk skor data PISA, sekolah sebagai pengaruh acak, memberikan kontribusi keragaman yang cukup besar. Hal ini juga serupa pada kajian empiris yang dilakukan pada rata-rata nilai UN SMA IPA di Jawa Barat. Penduga komponen ragam yang diperoleh menunjukkan variasi yang besar antar sekolah terhadap nilai rata-rata UN untuk keenam mata mata pelajaran.

Untuk menangani struktur data kompleks yang melibatkan peubah respon multivariat serta pengaruh acak dua level pada struktur data longitudinal, telah dikembangkan. Pengaruh acak dalam model diasumsikan mengikuti distribusi normal, dan waktu pengukuran tidak diasumsikan sebagai efek tetap maupun efek acak di dalam model. Estimasi parameter untuk pengaruh tetap dilakukan dengan menggunakan metode *Maximum Likelihood Estimation* (MLE), sedangkan estimasi komponen ragam menggunakan metode *Restricted Maximum Likelihood* (REML).

Performa dari MMLMM dilakukan dengan menggunakan sifat-sifat penduga berupa relatif bias dan *Root Mean Square Error of prediction* (RMSEp) yang dievaluasi melalui kajian simulasi. Hasilnya menunjukkan bahwa nilai bias relatif hampir tak berbias pada seluruh penduga tetap parameter. Bias relatif akan berkurang dan mendekati 0 seiring bertambahnya ukuran sampel. Nilai RMSEp juga semakin kecil seiring bertambahnya ukuran sampel. Dengan demikian, model yang diusulkan menunjukkan kinerja yang lebih baik dengan ukuran sampel yang besar. Hal ini menunjukkan bahwa usulan model *multilevel multivariate linear mixed model* (MMLMM) yang dibangun menghasilkan estimasi parameter tetap yang tidak bias, mempunyai varians minimum dan konsisten. Hasil studi simulasi, terlihat bahwa performa MMLMM lebih unggul dibandingkan MLMM.

Studi empiris selanjutnya dilakukan dengan penerapan MMLMM pada rata-rata nilai UN SMA jurusan IPA di Provinsi Jawa Barat. Hasilnya menunjukkan bahwa galat baku hampir seluruh penduga nilai parameter tetap dari analisis MMLMM lebih kecil dibandingkan dengan MLMM. Terungkap juga bahwa model yang diusulkan menghasilkan AIC dan BIC yang jauh lebih rendah jika dibandingkan dengan model umum. Hal ini menjadi bukti kuat bahwa performa MMLMM lebih unggul dibandingkan MLMM.

Dalam disertasi ini, kami menduga pengaruh tetap model secara terpisah dengan pendugaan parameter acak dan menurunkan formula pendugaan parameter secara analitik hanya untuk parameter tetap. Dalam penelitian masa depan, akan menjadi tantangan untuk mengembangkan algoritma komputasi dan simulasi dimana pendugaan pengaruh tetap dan acak dapat dilakukan secara serentak dan untuk menyelesaikan pendugaan komponen ragam secara analitik.

Kata kunci: data *cross-sectional*, data longitudinal, model campuran linear multiraviat, model campuran linear multiraviat multilevel, pengaruh acak dua-level.

SUMMARY

VERA MAYA SANTI. Enhancing Multivariate Linear Mixed Model with Two-Level Random Effects for Longitudinal Data and Its Application to PISA and National Examination Scores. Supervised by KHAIRIL ANWAR NOTODIPUTRO, INDAHWATI, and BAGUS SARTONO.

Increasing the breadth to involve multivariate responses in cross-sectional data analysis becomes essential when examining interconnected and multidimensional phenomena. Including multivariate responses enables researchers to examine relationships and dependencies among numerous variables concurrently. The necessity to analyze multivariate responses in cross-sectional data emerges when researchers seek to understand complex patterns, dependencies, and interactions among various factors within a specific population.

In actual scenarios, a series of observations is frequently taken multiple times, a phenomenon commonly known as repeated measurements. Repeated measurements entail evaluating a research subject at different observation points in time. Several researchers have devised alternative methods for analyzing longitudinal data. The linear mixed model is the more prevalent approach (LMM).

Another challenge in longitudinal data analysis arises from the presence of multivariate response variables, where multiple responses are observed in the survey. One approach to tackle this is by simultaneously modeling them using a multivariate linear mixed model. The complexity of real-world data sets further increases when random effects are of a multilevel or hierarchical nature. Substantial research questions emerging in longitudinal data structures, where the response variables are multivariate and the random effects are multilevel, often involve issues such as interrelated changes between response variables.

An example of hierarchical data with multiple observations is the PISA scores and national exam (UN) data for high school (SMA) students. This research is conducted repeatedly to observe the trend of average PISA scores of Indonesian students and the average national exam scores at the high school level. PISA score data is designed to include a representative sample from the target population, which is the result of a survey conducted by the OECD. There are three test subjects in the OECD survey: reading literacy, mathematics, and science. Meanwhile, the average SMA national exam scores data is sourced from the Basic Education Data. (DAPODIK).

Cross-sectional studies have also been conducted to analyze PISA score data. In this case, it is crucial to capture whether there is any association or relationship between students' PISA scores across the three test subjects and also the school as the random effect. Furthermore, longitudinal studies have also been conducted. This study is applied to the average national exam (UN) scores data for high school students majoring in science in the West Java province. In this longitudinal study, it is essential to capture trends among the average UN scores for six subjects: Indonesian language, English language, mathematics, physics, chemistry, and biology over time. Another important aspect is to ascertain whether schools contribute as a random effect to the response variables.

Empirical analysis using MLMM has been conducted for both cross-sectional and longitudinal data structures. Although PISA score data has been the subject of

several studies in Indonesia, simultaneous analysis of all three PISA score data using MLMM has not been done before. Similarly, simultaneous analysis of average UN scores for high school students majoring in science in West Java with longitudinal data structure and the addition of random effects using MLMM has not been conducted. The results indicate that with the addition of random effects in the model, it is possible to reduce the standard errors of the fixed parameter estimates. Additionally, for PISA score data, schools as random effects contribute significantly to the variance. This finding is consistent with empirical studies conducted on the average UN scores for high school students majoring in science in West Java. The estimates of the variance components obtained show significant variation between schools regarding the average UN scores for the six subjects.

To handle complex data structures involving multiple outcomes and two-level random effects in longitudinal data, MMLMM (Multivariate Multilevel Linear Mixed Models) has been developed. In this model, the random effects are assumed to follow a normal distribution, and the measurement time is neither considered a fixed effect nor a random effect. Parameter estimation for fixed effects is carried out using the Maximum Likelihood Estimation (MLE) method, while the estimation of variance components utilizes the Restricted Maximum Likelihood (REML) method.

The performance of MMLMM is assessed using the properties of the estimators, namely relative bias and Root Mean Square Error prediction (RMSEp), which are evaluated through simulation studies. The results indicate that the relative bias value produces only a slight bias in all of the parameter estimators. The relative bias will decrease and approach 0 as the sample size increases. The RMSEp value also becomes smaller as the sample size increases. Thus, the proposed model shows higher performance with large sample sizes. This indicates that the proposed model that was built produces fixed parameter estimates that are unbiased, have minimum variance and consistent. Based on the simulation study results, it is evident that the performance of MMLMM is superior to that of MLMM.

The subsequent empirical study is conducted by applying MMLMM to the average scores of national exams for senior high schools specializing in natural sciences in West Java Province. The results indicate that the standard error of nearly all estimated parameter values from the MMLMM analysis is smaller than that of MLMM. It is revealed that the proposed model resulted in much lower AIC and BIC when compared to the common model. This is strong evidence that MMLMM's performance is superior to MLMM.

In this dissertation, we have partially estimated the fixed effects of the model and solely conducted analytical parameter estimation for the fixed parameters. In future studies, there will be a challenge to develop computational algorithms and simulations that enable simultaneous estimation of both fixed and random effects, as well as completing the estimation of variance components analytically.

Keywords: cross-sectional data, longitudinal data, multiraviate linear mixed model, two-level random effect, multilevel multiraviate linear mixed model.

ENHANCING MULTIVARIATE LINEAR MIXED MODEL WITH TWO-LEVEL RANDOM EFFECTS FOR LONGITUDINAL DATA AND ITS APPLICATION TO PISA AND NATIONAL EXAMINATION SCORES

VERA MAYA SANTI

Dissertation
as one of the requirements for obtaining a doctoral degree
on Statistics and Data Science

STATISTICS AND DATA SCIENCE
FACULTY OF MATHEMATICS AND NATURAL SCIENCE
IPB UNIVERSITY
BOGOR
2024

External Examiners on Closed Examination:

1 Dr. Ir. I Made Sumertajaya, M.Si
2 Dr. Kusman Sadik, S.Si., M.Si

External Examiners on Promotion Examination:

1 Prof. Dr. Komarudin, M.Si (Rektor UNJ)
2 Dr. Kusman Sadik, S.Si., M.Si

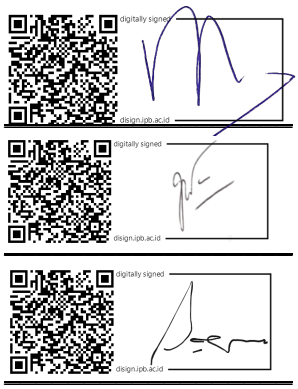
Dissertation Title : Enhancing Multivariate Linear Mixed Model with Two-Level Random Effects for Longitudinal Data and Its Application to PISA and National Examination Scores
Name : Vera Maya Santi
SIN : G161180021

Approved by

Chairman of Supervisory Commission:
Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S

Member of Supervisory Commission:
Dr. Ir. Indahwati, M.Si

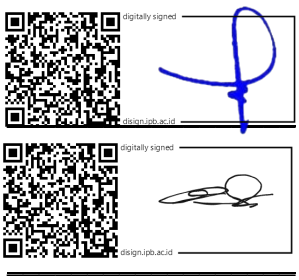
Member of Supervisory Commission:
Dr. Bagus Sartono, S.Si., M.Si



Acknowledged by

Head of Doctoral Study Program in Statistics and Data Science:
Dr. Anang Kurnia, M.Si
NIP 197308241997021001

Dean of the Faculty of Mathematics and Natural Sciences:
Dr. Berry Juliandi, S.Si., M.Si
NIP. 197807232007011001



Date of Examination:
(27th June 2024)

Date of Graduation:
(5th August 2024)

ACKNOWLEDGEMENTS

Alhamdulillaahirobbil'alamiin, I would like to thank ALLAH Subhanahu wa ta'ala, the most merciful and compassionate, for His support, help, and generosity, so that this dissertation is completed.

I acknowledge that the realization of this journey would not have been achievable without the support of my family, supervisors, and friends. Therefore, I express my heartfelt gratitude to those who have contributed of this dissertation and supported me in various ways throughout this remarkable journey. Their contributions were indispensable, and without their support, the completion of this research work would not have been possible. Moreover, in this occasion, I would like to express my gratitude to all parties that have played a part in the successful completion of this dissertation, including:

1. The supervisor committee, comprised of Prof. Dr. Ir. Khairil Anwar Notodiputro, MS., Dr. Ir. Indahwati, M.Si, and Dr. Bagus Sartono, S.Si, M.Si. You all have instilled in me the ability to think critically and develop into an independent researcher. I am profoundly grateful to all of you for your continuous encouragement, unforgettable help, and kind feeling throughout this study.
2. The Rector of Universitas Negeri Jakarta, the Dean of Faculty of Mathematics and Natural Sciences (FMIPA UNJ), and the Chairman of the Statistics Study Program have afforded the author the opportunity to pursue her doctoral program at IPB University.
3. The IPB graduate school and all staff who facilitated the author's study.
4. The Head of the Study Program and the lecturers in the Department of Statistics who have imparted extensive knowledge, enlightenment, and guidance throughout the years of study.
5. My beloved papaku Akhmad Mukhlis (Alm) and mamaku Nurhayati, for being such wonderful parents. Also, thank you to my mother-in-law Junizar Sofjan for your praying, support and love. Big love for my husband-Indra Octavia Kurnia and my lovely son Muhammad Rachmat Alifarizky, thank you for your blessing, praying, patience and love. Also my lovely sisters and brother, Leni Marlina, Akhmad Melvin Adriansyah and Charlia Venti Wahyuni. Your unconditional love and encouragement support me in finishing this work.
6. Doctoral students on Statistics IPB Class of 2018 and others who always encourage the author.

May this scholarly work be beneficial for those in need and contribute to the advancement of knowledge. The author apologizes for any shortcomings and errors in this dissertation.

Bogor, July 2024

Vera Maya Santi

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES