

Longitudinal survey data analysis

by

Pamela Opio Nasirumbi

Submitted in fulfilment of the academic
requirements for the degree of
Master of Science in Statistics,
School of Statistics and Actuarial sciences
University of KwaZulu-Natal

Pietermaritzburg

2006

Dedication

To my father, Mr Gabriel Opio and my mother, Dr Fina Asinansi Opio.

Declaration

The research work described in this thesis was carried out in the School of Statistics and Actuarial Sciences, University of Kwazulu-Natal, Pietermaritzburg, under the supervision of Dr. Temesgen Zewotir.

I, Pamela Opio Nasirumbi, declare that this thesis is my own, unaided work. It has not been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged.

Pamela

January, 2006

A large, stylized handwritten signature in black ink, appearing to be the name 'Pamela Opio Nasirumbi', written over a horizontal line.

Acknowledgement

I would like to thank Dr Temesgen Zewotir, my supervisor, for his direction, advice, patience and encouragement throughout the project. My deep appreciation goes to my lecturers Dr Henry Mwambi, Dr Peter Njuho, Mr Petrovius Horton and Mr Shaun Ramroop for their input, advice and encouragement. It has been a privilege to work with them.

I would like to thank the staff members and postgraduate students in the School of Statistics and Actuarial Sciences, and those in the School of Mathematics and Information Technology, University of KwaZulu-Natal, Pietermaritzburg for the hospitable environment they provided during my study. Special thanks go to my colleagues Gaetan Kabera and Winter Sinkala who were very helpful throughout my thesis write-up.

My special thanks also go to my friends who were always there to offer the support and encouragement which made my study possible. Special thanks go to Philip Awezaye who has been a pillar of support throughout the research.

I thank, with deep appreciation, the Center of Occupational and Environmental Health (COEH) department, University of KwaZulu-Natal Medical school, for providing me with all the data I required for the application of the principles in the thesis. Most especially, I thank Dr Rajen Naidoo, Dr Graciela Mentz and Dr MaryLou Thompson for all their assistance and input in the research. To the Almighty God, who makes every thing possible, I give thanks.

Abstract

To investigate the effect of environmental pollution on the health of children in the Durban South Industrial Basin (DSIB) due to its proximity to industrial activities, 233 children from five primary schools were considered. Three of these schools were located in the south of Durban while the other two were in the northern residential areas that were closer to industrial activities. Data collected included the participants' demographic, health, occupational, social and economic characteristics. In addition, environmental information was monitored throughout the study specifically, measurements on the levels of some ambient air pollutants. The objective of this thesis is to investigate which of these factors had an effect on the lung function of the children.

In order to achieve this objective, different sample survey data analysis techniques are investigated. This includes the design-based and model-based approaches. The nature of the survey data finally leads to the longitudinal mixed model approach.

The multicollinearity between the pollutant variables leads to the fitting of two separate models: one with the peak counts as the independent pollutant measures and the other with the 8-hour maximum moving average as the independent pollutant variables. In the selection of the fixed-effects structure, a scatter-plot smoother known as the loess fit is applied to the response variable individual profile plots.

The random effects and the residual effect are assumed to have different covariance structures. The unstructured (UN) covariance structure is used for the random effects, while using the Akaike information criterion (AIC), the compound symmetric (CS) covariance structure is selected to be appropriate for the residual effects. To check the model fit, the profiles of the fitted and observed values of the dependent variables are compared graphically. The data is also characterized by the problem of intermittent missingness. The type of missingness is investigated by applying a modified logistic

regression model missing at random (MAR) test.

The results indicate that school location, sex and weight are the significant factors for the children's respiratory conditions. More specifically, the children in schools located in the northern residential areas are found to have poor respiratory conditions as compared to those in the Durban-South schools. In addition, poor respiratory conditions are also identified for overweight children.

Contents

Dedication	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
Contents	ix
List of Tables	xi
List of Figures	xiii
1 Introduction	1
2 Durban-South health survey data	4
2.1 The sampling design and data collection	4
2.2 The variables	6
2.3 Data editing and cleaning	11

3	Sample survey analysis	14
3.1	Design-based approach to survey inference	16
3.2	Model-based approach to survey inference	17
3.2.1	The design-based and model-based approaches - How are they different?	19
3.3	Model-assisted design-based approach	24
3.4	Randomization-assisted model-based approach	25
3.5	The multistage sampling design	26
3.5.1	The mixed model	28
3.5.2	The hierarchical linear model	31
4	The longitudinal mixed model	35
4.1	The fixed-effects/mean structure	37
4.2	The covariance structure	40
4.3	Modelling the covariance structure	41
4.4	Model reduction	42
4.4.1	Tests for the significance of the fixed effects	42
4.4.2	Tests for the significance of the random effects	43
4.5	Estimation of the fixed and random effects	46
4.6	Test for normality	50
4.7	Inference for fixed and random effects	51
4.8	Longitudinal mixed model with missing observations	54
4.9	Missing data mechanisms	56

4.9.1	Missing completely at random (MCAR)	56
4.9.2	Missing at random (MAR)	56
4.9.3	Missing not at random (MNAR)	57
4.10	Fitting the longitudinal mixed model with missingness	60
4.10.1	The logistic regression model MAR test	61
4.10.2	The MAR tests for random dropouts	62
5	Identification of the model	64
5.1	Preliminary analysis	64
5.1.1	Classification of the height and weight measurements	64
5.1.2	The distribution of the children by the child-specific variables	68
5.1.3	The data structure of the SO ₂ pollutant measures	69
5.1.4	The fixed-effects structure	76
5.2	Selection of a preliminary mean and random-effects structure	79
6	Analysis of the peaks pollutant model	84
6.1	Exploring the covariance structure of the model	85
6.2	Model reduction	88
6.2.1	Tests for the significance of the random effects	88
6.2.2	Test for the significance of the fixed effects	90
6.3	Missingness	92
6.3.1	Applying the logistic regression model MAR test	93
6.4	Model checking	95

6.5	Inference from the peaks pollutant model	97
7	Analysis of the 8-hour maximum pollutant model	107
7.1	Model reduction of the model	109
7.2	Inference of the 8-hour maximum model	111
8	Conclusion	119
	References	123
	Appendices	136
A	Appendix tables	137

List of Tables

5.1	Malnutrition Classification Systems	67
5.2	The distribution of children by sex, asthma status and school	68
5.3	The frequency distribution of children by height, weight and school	69
5.4	Pearson Correlation coefficients and $\text{prob} > r $ under $H_0: \rho = 0$	72
5.5	The Variance inflation factor for the 10 pollutant independent variables	73
5.6	The eigenvalues of the correlation matrix for the 10 pollutant variables	74
5.7	The eigenvectors of the correlation matrix for the 10 pollutant variables	74
5.8	The rotated eigenvectors of the correlation matrix for the 10 pollutant variables	75
5.9	The Variance inflation factor for the pollutant measures with the peaks and the 8-hour maximum computed separately	76
6.1	Covariance structure criteria for the random effects of the peaks pollutant model	87
6.2	The four random-effects models for the peaks pollutant model	89
6.3	The likelihood ratio statistic with the correct, the naive asymptotic null distributions for comparing random-effects models and the comparison p-values for the peaks pollutant model	89

6.4	The likelihood ratio test statistics and p -values for the 2-way interactions, 2-way interactions with time, and 2-way interactions with squared time in the peaks pollutant model	91
6.5	The logistic regression model MAR test results for the response variable	95
6.6	The REML estimates, standard deviation and p -values for the peaks pollutant model	99
7.1	Covariance structure criteria for the random effects of the 8-hour maximum pollutant model	109
7.2	The four random-effects models for the 8-hour maximum pollutant model	110
7.3	The likelihood ratio statistic with the correct, the naive asymptotic null distributions for comparing random-effects models and the comparison p -values of the 8-hour maximum pollutant model	110
7.4	The likelihood ratio test statistics and p -values for the 2-way interactions, 2-way interactions with time, and 2-way interactions with squared time in the 8-hour maximum pollutant model	111
7.5	The REML estimates, standard deviation and p -values for the 8-Hour maximum pollutant model	114
A.1	The peak counts of the SO ₂ measurements on a particular date per school	137
A.2	The 8-hour maximum of the SO ₂ moving averages on a particular date per school .	138
A.3	The SO ₂ levels above the standard guideline of 191pb as well as the date, time and the school in which the high SO ₂ level occurs	139

List of Figures

5.1	The graphical representation of the peak counts.	70
5.2	The graphical representation of the 8-hour maximum of the SO ₂ moving averages.	70
5.3	Individual profiles of 30 randomly selected children in the study.	77
5.4	Mean profiles for the WVarFEV1 response variable.	77
5.5	The Loess smooth curves of WVarFEV1 against day for the random sample of respective asthma status sub-populations.	78
6.1	Plot of the variance of residuals against time generated from the peaks pollutant model.	86
6.2	Scatter plot of residuals by time.	87
6.3	Graphical representation of the proportion of respondents to WVarFEV1 by time.	92
6.4	The observed and predicted/fitted profiles of the WVarFEV1 values for the peaks pollutant model.	96
6.5	Histogram of the (a) EB estimates (for the random effects) and the (b) residuals for the peaks pollutant model.	97
6.6	Scatter plots of the (a) EB estimates and (b) residuals for the peaks pollutant model.	97
6.7	The peaks pollutant model fitted profiles of the WVarFEV1 values for children in Durban-South schools and those located in the northern residential areas.	104

6.8	The peaks pollutant model fitted profiles of the WDV _{Var} FEV ₁ values for the male and female children.	105
6.9	The peaks pollutant model fitted profiles of the WDV _{Var} FEV ₁ values for children of the three weight categories.	105
7.1	Plot of the variance of residuals against time generated from the 8-hour maximum pollutant model.	108
7.2	Scatter plot of residuals by time. The residuals are generated from the 8-hour maximum pollutant model.	109
7.3	The 8-hour maximum pollutant model observed and predicted/fitted profiles of the WDV _{Var} FEV ₁ values.	112
7.4	Histogram of the (a) EB estimates (for the random effects) and the (b) residuals for the 8-hour maximum pollutant model.	113
7.5	Scatter plot of the (a) EB estimates and (b) residuals for the 8-hour maximum pollutant model.	113
7.6	The 8-hour maximum pollutant model fitted profiles of the WDV _{Var} FEV ₁ values for children in Durban-South schools and those located in the northern residential areas.	116
7.7	The 8-hour maximum pollutant model fitted profiles of the WDV _{Var} FEV ₁ values for the male and female children.	117
7.8	The 8-hour maximum pollutant model fitted profiles of the WDV _{Var} FEV ₁ values for children of the three weight categories.	117

Chapter 1

Introduction

The Durban South Industrial Basin (DSIB) is one of the areas in South Africa characterized by a number of industries. It is located in the eastern coastal region of South Africa in the province of KwaZulu-Natal. Because of its geographic relationship with certain fixed sources of air pollutants, the DSIB is at a particularly high risk of exposure to significant ambient air pollution. Major emissions are from the so-called fugitive air emissions at refineries, from industrial and passenger vehicles, from other industries in close proximity as well as the Durban airport. The two major refineries in the basin are the petroleum refineries Engen and Sapref. There is also a pulp and paper manufacturer, Mondi. The recognition of this risk is reflected by an industry-funded South Durban Sulphur Dioxide Management Systems Committee (SDSDMSC) which, since June 2000, has been continuously monitoring one pollutant of concern, sulphur dioxide, at the Settlers Primary School. The available data on sulphur dioxide indicates that the average and/or maximum exposure at the Settlers school has frequently exceeded World Health Organization (WHO), the South African Department of Environmental Affairs (DEAT), and the SDSDMSC guidelines.

In light of this background, the Center of Occupational and Environmental Health

of the Nelson R. Mandela School of Medicine and the University of Michigan, USA were contracted to conduct a large-scale study to look at the effects of pollution on the health of exposed communities in the Durban-South industrial/residential basin, on behalf of the Durban Metro. The two year project, fully funded by Durban Metro, was part of a national cabinet decision to address the problems of pollution in this part of Durban.

The sole aim of this Durban-South health survey was to determine the health status of the Durban-South residents with specific reference to respiratory health outcomes and other chronic diseases. In addition, the project aim was to determine the relationship between environmental pollution, these health outcomes and the quality of life within this community, particularly among susceptible populations.

The study was conducted at seven primary schools namely: Nizam Road in Merebank, Assegai primary in Austerville, Dirkie Uys primary in Bluff, Entuthukweni primary at Lamontville township, Ferndale primary in Newlands, Briardale primary at Newlands West and Ngazana at KwaMashu. However, because of the vast amount of incompleteness of the data from the schools Entuthukweni and Briardale, only five schools were utilized in the thesis.

The focus of this thesis is to assess the relationship between the ambient air pollutant sulphur dioxide, the children's demographic and health characteristics, and their respiratory conditions. The major objective of this thesis is to determine which of the child demographic characteristics, child health characteristics and their exposure to the pollutant are the significant factors for the respiratory condition of the children. The children's respiratory condition is measured by the lung function measure known as the forced expiratory volume at one second (FEV1)

To achieve this objective, various sample survey analysis techniques will be explored,

more specifically the design-based and the model-based approaches. A choice will then be made of the most appropriate technique to accomplish the thesis objectives.

This thesis will be organized as follows. In Chapter 2, the nature of the Durban-South health survey data will be described in detail. This will also include the techniques applied to summarize the massive repeated measurements; and the data cleaning and editing methodologies applied to the data. Chapter 3 will review the sample survey data analysis techniques available specifically looking at the design-based and model-based approaches. The model-based approach is chosen which leads to the development of the longitudinal mixed model in Chapter 4. A review of the longitudinal mixed model with missing observations is also given in Chapter 4. The analysis and results of the Durban-South health survey will be dealt with in Chapter 5. Detailed inference of the data will be undertaken in Chapter 6 and Chapter 7. Finally, in Chapter 8, the findings of the thesis will be summarized.

Chapter 2

Durban-South health survey data

2.1 The sampling design and data collection

In the Durban-South health survey, seven communities were purposively considered for the study. Four were from Durban-South (Merebank, Lamontville, Wentworth/Austerville and Bluff) and three from the northern residential areas of the metropolitan boundaries (Newlands West, Newlands East and KwaMashu). The four Durban-South communities were considered due to their proximity to the source(s) of pollution. In addition, the communities from the two regions had similar characteristics, only differing in their status of industrialization, the Durban-South being more industrialized (the source of major environmental pollution) than the northern residential areas.

One primary school was purposively selected from each community. The selected schools were: Nizam Road in Merebank, Assegai primary in Wentworth/Austerville, Dirkie Uys primary in Bluff, Entuthukweni primary at Lamontville township, Ferndale primary in Newlands, Briardale primary at Newlands West and Ngazana at KwaMashu. The choice of the school was also based on its proximity to the pollution source(s) as

well as the proximity of the residences of the school children to the pollution source(s) under consideration. However, only five of these seven primary schools are used in this thesis as will be explained in Section 2.3.

From each school, one or two grade four classes were randomly selected and all children in this sample were included in the sample. An average of 70 children were expected from each school, therefore the second classroom was selected only if the first selected classroom was found to have a small number of children. In addition, all children with severe cases of asthma from grades 3 and 5 and the unselected grade 4 classes were also included in the sample. However the targeted number of 70 children per school was not met for Assegai, Dirkie Uys and Ngazana schools who had an original sample of 63, 51 and 67 children respectively. A total of 81 and 97 children were selected in the original sample from Nizam Road and Ferndale schools.

However, before any test could be performed or interviews conducted, informed consent was requested from the children's parents or guardians. It was made clear to all participants and their families that they were free to withdraw from the study at any time or refuse to participate in any aspect of the study without penalty. Informed consent included explanation of the expected benefits and risks of participation in each aspect of the study. Consent forms were therefore issued to the parents or guardians of the children in the sample and only those that offered their consent participated in the study. The percentage of the consented children in randomly selected grade 4 classrooms in Nizam Road, Assegai, Dirkie Uys, Ferndale and Ngazana schools were 72.31%, 70.45%, 34.78%, 68.67% and 81.03% respectively. Among the additional children with severe cases of asthma in grade 3, 5 and the unselected grade 4 classrooms, only 72.0%, 38.89% and 50.0% respectively offered consent to participate in the study.

The variables of interest in this thesis include the child forced expiratory volume at one second (FEV1), school location, sex, asthma status, height and weight and the

sulphur dioxide (SO₂) pollutant measures. The main focus of this thesis is to determine which of the child characteristics namely school location, sex, asthma status, height and weight have an effect on the respiratory conditions of the children measured by the FEV1. Also to be investigated is whether the daily fluctuations in the ambient air pollutant SO₂ measured at the schools are predictive of the fluctuations in pulmonary function measures that depict the respiratory condition of the child.

2.2 The variables

The response/dependent variable is the lung function measure FEV1. These repeated measurements are bihourly measures of the pulmonary function during the school day and they are aimed at measuring the respiratory condition of the child. They were obtained by use of a machine commonly known as an airwatch instrument. These measurements were obtained in four intensive phases, each phase being a period of three weeks. The thesis considers only the first of the four phases which spanned a total of fourteen days. The one missing day of the third week was a public holiday and since all measurements were made only at the school, no measurements were made on that holiday.

In the three weeks, on each of the five school days during the week, the children performed airwatch maneuvers every 1½ to 2 hours (that is four times in the 5.5 hour school day which is approximately at 08h00, 09h45, 11h30 and 13h20). At each blow time, the child blew five times into the machine making a total of twenty blows or readings for a child per day. For analysis purposes, this information is summarized into one reading a day per child utilizing one of the summary methods used in the Settlers primary school health study (Robins *et al.*, 2002). This is the within-day

variability (WVarFEV1) calculated using the formula

$$\frac{\text{Best reading} - \text{Worst reading}}{\text{Best reading}} \quad (2.1)$$

The best reading is the largest (maximum) of the day and the worst reading is the lowest (minimum) of the day. This reduces the respiratory measurements to a maximum of fourteen repeated measurements per child.

The Settlers primary school health study was a Durban-based project with similar objectives to the Durban-South community project but was conducted on a much smaller scale. It was from the findings of the Settlers project that the decision to carry out the Durban-South project was made.

Early research has shown that increased ambient air pollution levels, particularly of SO₂ (Abbey *et al.*, 1993; Katsouyanni *et al.*, 1997) precipitate symptoms of asthma and increase emergency department visits and hospitalizations among children and adults with asthma and other chronic respiratory conditions (Schwartz *et al.*, 1993a; Walters *et al.*, 1994). Highly sensitive sub-populations such as asthmatics, the elderly and children are at increased risk. Exposure to ambient particulate matter (PM) and co-pollutants such as sulphur dioxide in the ambient environment may provide the critical factor in increased morbidity and mortality in these individuals in urban centers (United States Environmental Protection Agency, 1996; Schwartz, 1993b). Acid aerosol is closely associated with SO₂ in the eastern U.S. and has been associated with increased respiratory hospitalizations (Thurston *et al.*, 1994).

Seven key ambient air pollutant exposures were investigated in the Durban-South health survey including the air pollutants sulphur dioxide (SO₂) and ambient particulate matter (PM). The PM data was not readily available so only the SO₂ data was considered in this thesis. The SO₂ is the result of the combustion of fuels containing sulphur for example, coal and oil, gasoline production and oil refining, and smelting

ores.

Monitors to register the SO₂ readings were placed at each of the schools. Two assumptions were made in this case, one being that the children spent most of their day at school and the other being that the child resided close to their specific schools. Therefore the monitors situated at the schools registered their average exposure to the SO₂ pollutant. These SO₂ monitors recorded readings every 10 minutes, measured in parts per billion(ppb). A reading of 100 ppb means that 0.00001 percent of the air around the monitor is SO₂ gas. For analysis purposes, this per 10-minute data was summarized into a per day reading utilizing the summary criteria used in the Settlers primary school health study (Robins *et al.*, 2002). The summary measures were the peak counts and the 8-hour maximum of the moving averages.

To compute the peak counts, the pollutant measures obtained are compared with the national standard guidelines provided by the modern air quality monitoring network established in the Durban-South basin. These guidelines were designed to be protective of public health and are displayed by the modern air quality monitoring network in the weekly reports in the website www2.nilu.no/airquality/. From their most recent report (Report summary week 24, 2005 running from 13/June/2005 to 19/June/2005), the per 10-minute SO₂ standard reading was 191 ppb. The per 10-minute SO₂ readings were then compared with the 10-minute standard of 191 ppb to determine whether it was above or below the 191 ppb mark. Therefore a peak was any per 10-minute SO₂ reading above the per 10-minute standard of 191 ppb.

In the calculation of the peak counts, the number of times the pollutant levels were above the standard guidelines were counted. This was computed to correspond to the 24 hours prior to the first airwatch blow. For example, the peaks corresponding to date 6/14/2004 are the total number of times the pollutant levels went above the standard guidelines between 8am of 6/13/2004 and 7.55am of 6/14/2005. These peak counts are

referred to as 'peaks' throughout the thesis.

In the Settlers primary school health study (Robins *et al.*, 2002), possible lag effects (hours to days) were modelled to account for possible prior exposure effects. Examined models considered the effects of the lung function to exposures occurring earlier the same day (lag 0), the previous day (lag 1), and two days prior (lag 2). Prior one day and prior two days exposure to SO₂ were found to be associated with highly statistically significant increases in the occurrence for lower respiratory symptoms. Due to this outcome, it seemed best to investigate the lag effects from the previous day (lag 1), two days prior (lag 2) to five days prior (lag 5) in this thesis. Thus, an additional four other pollutant variables were considered and these were the peak counts two days, three days, four days and five days prior to the day of the FEV1 measurement.

The alternative summary pollutant measure was obtained by selecting the maximum of the 8-hour moving averages computed 24 hours before the first blow of the day. For example the 8-hour maximum measure corresponding to the 31st of May 2004 is calculated as follows: First attained is the average of the 10-minute readings from 0 hours to 7:50 hours of the 30th of May 2004. There after 8-hour moving averages are computed for the next 24 hours that is up to 7:50 hours of the 31st May. To correspond to the time of the first airwatch blow of the 31st of May, which was at 8:00 hours, the maximum of the moving averages in the 24 hours is attained to serve as the 8-hour maximum single record for that date. These values are referred to as '8-hour maximum' throughout the thesis. Still taking the lag effects into consideration, four other measures, which are the 8-hour maximum pollutant measures two days, three days, four days and five days prior to the day of the FEV1 measurement, were considered.

Unlike the peaks pollutant measures that only represent the number of times the pollution measures go beyond the acceptable standards, the 8-hour maximum pollutant measures take into consideration the actual pollutant measurements averaging them

appropriately into one measure.

Previous studies have shown that the taller the person is, the greater his/her lung volume (Buist, 1982; Buist and Vollmer, 1988). In addition the lung function decreased at both extremes of weight (Schoenberg, Beck and Bouhuys, 1978; Dockery *et al.*, 1985). After correcting for body size, girls appear to have higher expiratory flows than do boys, whereas adult men have larger volumes and flows than women (Buist, 1982; Schwartz, Katz, Fegley and Tockman, 1988a; 1988b). Since the lung function of the children was under consideration, it seemed appropriate to include the variables: child height, weight and sex in the analysis.

The child height and weight data was obtained during the baseline spirometric assessment and methacholine challenge tests that were administered to all the children in the child sample. These lung function tests were done using a spirometry instrument. From questionnaire instruments administered to the children and their guardians, information on the age and sex of the children was obtained. Using the asthma status classification criteria developed by the medical specialists, the asthma status of the child was determined from the child respiratory symptom responses in the questionnaires. The asthma status was categorized as being severe or, mild or normal. Children with severe cases were referred to as 'persistent asthmatics', those with a mild case of asthma as 'asthmatics' and those with no cases of asthma as 'normal'.

In conclusion, the variables in the data set have the following structure: variable *Child* as the random factor; variables *School*, *Asthma status*, *Age*, *Height*, *Weight* and the 14 repeated SO₂ pollutant readings (the Peaks and the 8-hour maximum) as fixed factors; and the 14 repeated FEV1 measurements per child as the response variable. The data is therefore longitudinal in nature since the pollutant readings and the FEV1 measurements were sequentially measured over time.

2.3 Data editing and cleaning

In any survey, non-sampling errors cannot be avoided but their occurrence can be minimized. These errors occur at the data collection and processing stages of the survey operation (Dalenius, 1988; Hansen, Hurwitz and Madow, 1953; Thompson, 1992; Lessler and Kalsbeek, 1992; and Kish, 1965). These errors have to be identified and dealt with before the analysis of the data. To ensure the collection of adequate data, all who were to be involved in the study went through a period of intense training in the areas in which they were going to participate.

The FEV1 measurements were obtained by children blowing into instruments known as airwatches which in turn produced readings that were later downloaded for analysis. Since the airwatch maneuvers were carried out at schools, both the children and their teachers were trained on the proper usage of the airwatch devices prior to the blow exercises. They were also trained on the proper handling and storage of the airwatch. The airwatches were also properly calibrated and adjusted appropriately before being handed over to the children. Each child had his/her airwatch labelled by name and serial number.

Despite the intensive training received by both the respondents and the interviewers, an experience from the Settlers primary school health study (Robins *et al.*, 2002) was utilized to detect invalid blows. In this procedure, histograms of the FEV1 values were plotted and extreme values truncated. The values that remained were compared with the baseline FEV1 of the specific children. It was discovered that on average the remaining FEV1 values lay between 30% and 120% of the baseline value. As a result all those outside the range were considered invalid FEV1 readings. The aim was to have a symmetrical confidence interval of the FEV1 readings. The baseline FEV1 on the other-hand was computed during the lung function tests using prediction equations

of the race, sex, age and height. The aim of these equations was to predict what the reading of a specific child would be if he/she was normal.

I spent a great amount time on the editing and cleaning of the data. Since every child was expected to blow 20 times a day for 14 days, a total of 280 records were expected per child. Each of these records were checked to make sure that they corresponded to the correct time and date of the blow. Some data was found to correspond to dates and times outside the actual dates and times of blows. Cleaning of the times and dates was done by use of the check-lists kept by the interviewers during the FEV1 blow exercises. However, there were times when the child did not have any airwatch reading or when an invalid reading was recorded. Such data was considered as missing data.

The SO₂ monitors, situated at each of the schools, were calibrated at the beginning of the study. Quality checks were then performed approximately weekly and at the end of the study. During the regular monitor calibrations and data downloads, abrupt hikes or falls in the readings were noted and appropriate adjustments were made. These extreme readings were commonly due to the malfunction of the monitors. Spot-on repairs were therefore also made in such instances. Problems with calibrations were reported especially at the Entuthukweni and Briardale schools at the beginning of the study but were successfully sorted out later in the study. This however led to a vast amount of incompleteness in the data relating to the first intensive phase of the study known as phase 1. It is for this reason that data from these two schools is not considered in this thesis. However the other five schools also had a few cases of missing data. This meant that the handling of missing data needed to be included in the decision on the appropriate analysis technique chosen for the study.

The child characteristics of age and sex were obtained from the questionnaires administered by trained interviewers. To minimize non-response, interviewers made

appointments with the school head teacher and teachers prior to the interviews. During the initial visits to the schools the interviewers informed the teaching staff about the advantages of the survey. They were encouraged to assist in the exercise as well as to encourage the parents or guardians of the children to participate in the study.

Before data capture, all the questionnaires were carefully edited by trained project staff to ensure that all the required information was obtained by the interviewers. This information was then double-entered to minimize data entry errors. Double-entry of data is a technique in which data is captured twice with the second data entry being a form of check for any errors in the initial data entry. More cleaning was done on the captured data which in some cases resulted in some call-backs.

Chapter 3

Sample survey analysis

In any statistical investigation, one has to decide whether to cover the whole population, or carry out the statistical investigation based on a sample. A complete enumeration of the whole population is known as a census. On the other hand a statistical inquiry based on a sample on which inferences are made about the population is known as a sample survey. In many fields of scientific study, most of the data is collected using sample surveys usually due to their advantages of being, quick, timely, manageable to carry out and less costly. The aims of a sample survey can be broadly classified as: to make inferences about the population parameter(s) of interest, and to investigate the relationships among the variates in the study.

Over the years, numerous sampling schemes in sample survey have emerged and some of those in existence have been modified. For instance, in 1925 the use of simple random sampling in official statistics was accepted mainly because it obtained representative samples but also because of the existence of the theory of measuring the uncertainty due to sampling (Smith, 1999). However, the difficulty of extending this theory to more complex sampling schemes led to the next major advancement in the sample survey methods by Neyman (1934). He changed the theoretical basis of sample

survey inferences from being based on the hypergeometric likelihood function to being based on sampling errors with a randomization distribution. The latter is what was known as the design-based approach to survey inference. In this approach a researcher draws a probability sample from a finite population for making inferences about this finite population. The probability model for the sample data depends on how the sample was drawn by the researcher (Snijders and Bosker, 1999). Procedures which properly account for the design and population structure are then developed to make inferences about the parameters. (Skinner, Holt and Smith, 1989). Therefore, this approach is based on the randomization distribution induced by the survey design (Levin, 1999).

In the 1950s and 1960s, theoreticians and scholars addressed the foundations of this design-based inference approach with a number of them largely holding negative views about the theory (Smith, 1999). This led to a search for an alternative model-based framework of survey inferences. In this approach, inferences are extended to include not only the survey variables of interest as explanatory variables but also the variables used in the survey design (Skinner *et al.*, 1989). Thus, inferences are made using the best model that explains all structural and stochastic variations in the sample data. The approach assumes a probability model for the sample data and inferences are about some large and hypothetical super-population. The main aim of developing this model-based inference approach was to enable the investigation of the relationships between variables in the survey population. This was done by taking the finite population to be a realization of an infinite super-population (Levin, 1999). A role is therefore seen for models in descriptive as well as analytic surveys.

3.1 Design-based approach to survey inference

The design-based approach to survey inference is described in many texts for example Hansen, Hurwitz and Madow (1953), Kish (1965), Cochran (1977), and Little (2004). The following description by Little (2004) is not completely general but captures the main features.

For a population with N units, let the random vector Y be given by $Y = (y_1, \dots, y_N)$ where y_i is the set of survey variables for unit i . Also let $I = (I_1, \dots, I_N)$ denote the set of inclusion indicator variables where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ if it is not included. Let Y_{inc} be the observed part of Y and Y_{exc} the excluded part of Y . The design-based inference is based on the distribution of I with the survey variables Y treated as fixed quantities. Inference about the finite population quantity say, $Q = Q(Y)$ involves the following steps:

- (a) the choice of an estimator say, $\hat{q} = \hat{q}(Y_{inc}, I)$. This is a function of the observed part of Y that is unbiased or approximately unbiased for Q with respect to the distribution of I . The estimator \hat{q} is a random variable as a function of I , not Y_{inc} which are fixed quantities.
- (b) the choice of a variance estimator $\hat{v} = \hat{v}(Y_{inc}, I)$ that is unbiased or approximately unbiased for the variance of \hat{q} with respect to the distribution of I . Inferences are then generally based on normal large sample approximation. For example, a 95% confidence interval for Q is $\hat{q} \pm 1.96 \sqrt{\hat{v}}$.

To describe the above process for a stratified random sample, Little (2004) considers a simple case of estimating a finite population mean \bar{Y} . Suppose that the population is divided into K strata, and that N_i is the known population count in stratum i while \bar{Y}_i is the unknown population mean of stratum i . Then the quantity of interest in this

case is

$$Q = \bar{Y} = \sum_{i=1}^K P_i \bar{Y}_i,$$

where $P_i = N_i/N$ is the proportion of the population in stratum i . Assume that a random sample of size n_i of N_i units is taken from stratum i and $y_{ij} : i=1, \dots, K; j=1, \dots, n_i$ denotes a set of sampled Y -values from stratum i , then $Y_{inc} = \{y_{ij} : i=1, \dots, K, j=1, \dots, n_i\}$. Stratified random sampling has the property that all possible samples of size n_i from stratum i have the same probability of being selected which is given by

$$P(I_{ij} = 1) = \begin{cases} \frac{1}{\binom{N_i}{n_i}} & \text{if } \sum_{j=1}^{n_i} I_{ij} = n_i \\ 0 & \text{otherwise.} \end{cases}$$

The usual estimator of \bar{Y} in this setting is the stratified mean

$$\hat{q} = \bar{y}_{st} = \sum_{i=1}^K P_i \bar{y}_i = \sum_{i=1}^K \frac{N_i}{N} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^K \frac{1}{\pi_i} \frac{1}{N} \sum_{j=1}^{n_i} y_{ij}.$$

The estimated variance of the stratified mean is

$$\hat{v}_{st} = \sum_{i=1}^K P_i^2 S_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right), \quad (3.1)$$

where S_i^2 is the sample variance for stratum i . The quantities \bar{y}_{st} and \hat{v}_{st} , given by (3.1) and (3.1), are used for constructing the 95% confidence interval for \bar{Y} of the form $\bar{y}_{st} \pm 1.96 \sqrt{\hat{v}_{st}}$ and for the hypotheses tests about the population mean \bar{Y} .

3.2 Model-based approach to survey inference

The model-based approach to sample survey inference, on the other hand, requires a model for the survey outcomes Y_{inc} . This model is then used to predict the non-sampled values, Y_{exc} , of the population and hence the finite population quantities, Q . Little (2004) divides the model-based approach into super-population modelling and Bayesian modelling. In super-population modelling (for example Royall 1970; and

Valliant, Dorfman and Royall, 2000), the population values of Y are assumed to be a random sample from a super-population with a probability distribution $p(Y/\theta)$ indexed by a fixed parameter vector θ . Inferences are based on the joint distribution of Y and the inclusion indicator variables I . The Bayesian modelling approach, on the other hand, requires the specification of a prior distribution $p(Y)$ of the population values (Ericson, 1969; Binder, 1982; Rubin, 1983, 1987; Ghosh and Meeden, 1997). Inferences about the finite population quantities $Q(Y)$ are then based on the posterior predictive distribution $P(Y_{exc}/Y_{inc})$ which is the distribution of non-sampled values (Y_{exc}) of Y given the sampled values (Y_{inc}). The specification of the prior distribution $p(Y)$ is often achieved via a parametric model $P(Y/\theta)$ indexed by parameters θ combined with prior distribution $p(\theta)$ for θ .

The specification of $p(Y/\theta)$ in this Bayesian formulation is the same as in parametric super-population modelling. In large samples, the likelihood based on this distribution dominates the contribution from the prior of θ . As a result, large-sample inferences from the super-population and Bayesian modelling approaches are often similar. Model formulations thus do not involve the distribution of I , but the distribution of Y alone.

A key motivation for probability sampling from the modelling perspective is that it avoids the need to specify a model for the sampling mechanism even though the sampling distribution is not the basis for inference. From the Bayesian perspective, random sampling provides a justification for assumptions of exchangeability of the sampling units that underpin identically and independently distributed (*iid*) models.

To describe the model-based inference for the mean from a stratified random sample, Little (2004) lets y_{ij} denote the value of Y for unit j in stratum i . A common baseline for continuous outcomes assumes that y_{ij} is normally distributed with mean μ_i and variance σ_i^2 . A simple Bayesian specification in the absence of strong prior knowledge

about μ_i and σ_i^2 yields a posterior mean that equals the stratified mean from the design-based inference. For example, when σ_i^2 is replaced by S_i^2 , the posterior variance is the same as the design-based variance. Thus for large samples, the posterior distribution of \bar{Y} yields a 95% posterior probability interval that is the same as the design-based 95% confidence interval. The former is interpreted as a probability statement about the unknown population mean and the latter as a confidence interval for the unknown population mean.

3.2.1 The design-based and model-based approaches - How are they different?

Different frameworks or approaches for inference depend on the assumptions about the processes which generate the sample data (Smith, 2001). The decision to use the model-based inference is a statement that the sample selection mechanism can be ignored for inference. Likewise, the decision to use the design-based inference is a statement that models have no relevance to the inferential framework.

According to Little (2004), the design-based inference takes into account features of the survey design, and provides reliable inferences in large samples without the need for strong modelling assumptions. On the other hand this inference approach yields limited guidance for sample adjustments. For example, the stratified mean which sampled units by the inverse of their probability of selection. The Horvitz-Thompson (HT) estimator applies this idea more generally as follows. Consider inference about the population total $Q(Y) = T = Y_1 + \dots + Y_N$ and a sample survey design with a positive inclusion probability $\pi_i = E(I_i/Y) > 0$, $i = 1, \dots, N$. Then the HT estimator is

$$\hat{t}_{HT} = \sum_{i \text{ sampled}} \frac{Y_i}{\pi_i} = \sum_{i=1}^N \frac{I_i Y_i}{\pi_i}$$

and is design unbiased for T since

$$E[\widehat{t}_{HT}/Y] = \sum_{i=1}^N E(I_i/Y) \frac{Y_i}{\pi_i} = \sum_{i=1}^N \pi_i \frac{Y_i}{\pi_i} = \sum_{i=1}^N Y_i = T.$$

The unbiasedness of \widehat{t}_{HT} under very mild conditions conveys robustness to modelling assumptions and makes it a mainstay/foundation of the design-based approach. However, the HT estimator has two major deficiencies. First, the choice of variance estimator is problematic for some probability designs such as systematic sampling. Second, the HT estimator can have a high variance, for example, when an outlier in the sample has a low selection probability and hence receives a large weight. On the other hand, independent and identically distributed (*iid*) models fail to account for sample survey design (Little 2004).

The differences between the design-based and model-based inference approaches to statistical analysis arise early in the study of statistics, in the form of the role of weights in multiple regression (Little 2004). For example, the basic fitting algorithm for standard forms of normal linear regression is ordinary least squares (OLS). OLS is based on a model that assumes a constant residual variance for all values of the covariates. However, if the variance of the residual for unit i is σ^2/u_i for some known constant u_i , then better inferences are obtained by weighted least squares (WLS) with unit i weighted proportional to u_i . This form of weighting is model-based since the linear regression model has been modified to incorporate a non-constant residual variance.

A quite different form of weighting arises in survey sampling, based on the sample selection probabilities. If unit i is sampled with selection probability π_i , then the OLS is replaced by the WLS by weighting the contribution of unit i to the least square equations by the inverse of the probability of selection ($w_i \propto \frac{1}{\pi_i}$). This form of weighting is design-based, with π_i , relating to the selection of units. Since unit i represents $1/\pi_i$ units of the population, it receives a weight proportional to $1/\pi_i$, in the regression.

The model-based approach (under the appropriate super-population probability models) also differs from the conventional design-based one in that it focuses on the characteristics of particular samples rather than on plans for choosing samples (Royall and Herson, 1973). The correctness of the results in the model-based approach depends on the super-population probability model whose perfect confidence in it is never justified.

By paying close attention on the asymptotics, Kott (2002) improved a bit on the variance estimator concentrating first on those situations where finite population correction matter and then on cases where the sample itself is not very large. As the population and then the sample become less large, it was necessary to make more assumptions about the error structure of the model. Relying on randomization-based properties, which are asymptotic in nature makes little sense. Models can therefore help in ferreting out just when the sample may be too small to assert asymptotic normality to the pivotal $(t_R/V$ when finite population correction is ignorable. Parameter t_R represents the regression estimator; V is the model variance estimator of t_R), a generally common and often an unjustified practice.

The use of a particular model-based approach helps to provide guidance in the choice of sampling designs but it is difficult to determine how far it should dictate the choice of an estimation (Hedayat and Sinha, 1991). Also, any deviation from the true model would make it difficult to validate the survey estimates unless the estimators were robust in the appropriate sense.

Kalton (1983) and Thomsen and Tesfu (1988), who have studied the use of the model-based approach in sample surveys, argue that although it is true that most sample surveys are aimed at providing estimates for simple descriptive parameters (like mean, variance) of the population of interest, typical sample surveys are multi-purpose in nature and different parts of the multi-response survey data may need to be

analysed for different purposes. Even if a model is suitable for one survey variable, it may not be realistic for others. Therefore Kalton (1983) and Thomsen and Tesfu (1988) conclude that a complete reliance on a particular super-population is not advisable in a multi-response multi-purpose survey but may be used as a guiding factor for selecting a reasonable sampling strategy.

The super-population parameters in the model-based approach may often be preferred to finite population parameters from the design-based approach as targets for inference in analytic surveys (Skinner *et al.*, 1989). However, if the population size is large, there will often be little numerical difference between the two approaches. A difference will therefore only be expected as the population size gets smaller. On the other hand, samplers often mistrust the model-based approach to survey inferences when applied to large samples from complex populations due to the concerns about model misspecifications (Zheng and Little 2003). However, the approach can work very successfully in survey settings provided the chosen models take into account the sample design and avoid parametric assumptions thus in favor of the design-based approach.

The model-based inferences are well equipped to handle complex design features such as cluster sampling through random cluster models, stratification through covariates that distinguish strata, non-response (Little, 1982; Rubin, 1987; Little and Rubin, 2002) and response errors. The model-based approach also allows for prior information to be incorporated when appropriate (Little, 2004).

Another area where the model-based methods are widely used is in small domain estimation. This is because small sample sizes can be too small for the design-based approach to obtain reliable inferences. However, aggregating the small domains would surely result in the results being based on a large enough sample, thus enabling the design-based principles to offer some protection against model failure. Despite the fact that the model-based approach gives powerful inferences, grossly invalid inferences

would result if the assumed model were unrealistic (Cassel, Särndal and Wretman, 1977).

But, all models are simplifications and are hence subject to some degree of misspecification. The major weakness of the model-based inference, therefore, is that if the model is seriously misspecified, it can yield inferences that are worse than the design-based inferences. A model for stratified sampling that ignores stratum effects is too misspecified for it to be a reliable basis for inference unless there are convincing reasons to believe that stratum effects are not present.

The role of randomization in experimentation is also an important and essential one. Its basic function is that it reduces the number of assumptions required for the validity of statistical inference procedures (Royall, 1970). This function has many aspects and the three frequently described are: to protect against failure of certain probabilistic assumptions; to average out effects of unobserved or unknown random variables; and to guard against unconscious bias on the part of the experimenter. All these objectives contribute to the general goal of making the results of the experiment convincing to others. Two other arguments that advanced in support of randomization in sampling problems were: that it enabled the sampler to check the accuracy of the assumptions concerning the relation between the response and explanatory variables; and also enabled the sampler to estimate the variance of the estimate accurately.

In practice, many survey statisticians, depending on the situation, have adopted both design-based and model-based philosophies of statistical analysis. For example, descriptive inferences about finite population quantities based on large probability samples are carried out using design-based methods but models are used for problems where this approach does not work such as non-response and small area estimation. This approach has increased in popularity, but as Little (2004) remarked, "one should not be satisfied with two competing statistical theories". This has led to the development of

approaches such as the model-assisted design-based approach and the randomization-assisted model-based approach.

3.3 Model-assisted design-based approach

In the model-assisted design-based approach, the design-based inference is treated as the real goal of the survey sampling, but models are employed to help choose between valid design-based alternatives. This leads to unbiased models and a small model-expected randomization mean square error. Therefore, the super-population models are not the basis for inference in the design-based approach, but are useful in motivating the choice of estimator (Little, 2004).

In particular, many of the classic estimators, such as the ratio estimator and the regression estimator, which incorporate covariate information, can be motivated as arising from linear super-population models. For example, the HT estimator can be regarded as a model-based estimator for the following linear model relating y_i to π_i :

$$y_i = \beta\pi_i + \pi_i\varepsilon_i,$$

or equivalently

$$z_i = y_i/\pi_i = \beta\varepsilon_i.$$

where ε_i are assumed to be iid normally distributed with mean zero and variance σ^2 . These models lead to $\hat{\beta} = n^{-1} \sum_{i \in S} y_i/\pi_i = t_{HT}/n$ where n is the sample size. The corresponding prediction for unit i is $\hat{y}_i = \hat{\beta}\pi_i$, and the prediction estimator of the total is thus

$$\hat{T}_{pred} = \sum_{i=1}^N \hat{y}_i + \sum_{i \in S} (y_i - \hat{y}_i) = \hat{t}_{HT} + \sum_{i \in S} (y_i - \hat{y}_i)$$

This estimator differs from the design-based HT estimator mentioned in Section 3.2.1 by a quantity that tends to zero with the sampling fraction n/N . This analysis suggests

that the HT estimator is likely to be a good estimator if the linear model relating y_i to π_i adequately describes the population. However, this estimate will surely be inefficient if the linear model does not fit the population.

3.4 Randomization-assisted model-based approach

Unlike the model-assisted approach, the randomization-assisted model-based approach treats the model-based inference as the goal of the survey sampling. However, design-based methods are employed to protect against inevitable model failures. However, the choices for the estimation strategy and the variance estimator between the model-assisted and the randomization-assisted approaches do not change in typically large samples and population environments. In this approach the estimators are also design-consistent.

Restricting attention to design-consistent estimation strategies assures that even when the model fails, the estimator will likely not be too far off from what it is estimating. Also for an estimation strategy (an estimator coupled with a sampling design) to be design-consistent, its mean square error should approach zero as the sample size becomes arbitrarily large. The emphasis of the randomization-assisted model-based approach is thus to make the bias small, thus leading to superior coverage estimates. In his paper, Kott(2002) concludes that although the dominant design-based and model-assisted design-based approaches have been fruitful in many ways, they should be supplanted by the randomization-assisted model-based approach. This is because inference should be based on the actually observed sample rather than on averaging over all potential samples. He warned that although the design-based methods provide some protection against model failure, their protection relies on invoking the asymptotic properties in a finite population. Thus, he argues that the model-assisted

routine actually makes more sense under the randomization-assisted model-based approach. However, the approach to handling multi-phase sampling designs from the randomization-assisted model-based viewpoint still needs to be developed.

In conclusion, one can say that the model-assisted approach takes the model out of the shadows of design-based survey sampling and gives it a formal place in survey theory and practice. However, the real inference is deemed to be related to the sampling design and not to the model distribution. This is done by choosing among design-consistent estimation strategies, by hypothesizing a reasonable and practical model, restricting attention to model-unbiased estimators, and by selecting that strategy with the minimum model-expected randomization mean square error. This routine however, makes even more sense under the randomization-assisted model-based approach.

3.5 The multistage sampling design

Most of the sources of variability in survey data are nested. Observations on children in classes, employees in firms, suspects tried by judges in courts, animals in litters, and longitudinal measurements on subjects are some of the examples of nested data structures (Snijders and Bosker, 1999). Data structures of this nature are viewed as multistage (or multilevel) samples from clustered populations.

The model-based design allows the researcher to go beyond estimating the summary measures into the casual explanation of the processes that underlie the descriptive survey approach in the analysis of complex survey data (Skinner *et al*, 1989 Chapter 1). This makes the model-based design appropriate approach for a multistage sampling design. In addition, this modelling approach can be adopted to complex finite population structures and sampling schemes (Ericson, 1969; Scott and Smith, 1969; and Royall, 1970). Complexity in this context is defined by the degree of complexity of the

sampling design.

Longitudinal data in particular are often highly unbalanced in the sense that the number of measurements per subject are unequal and/or that measurements are not taken at fixed time points. Due to their unbalanced nature, many longitudinal data sets cannot be analysed using multivariate regression techniques. The natural alternative arises from observing that subject-specific longitudinal profiles can often be approximated by using linear regression functions. In this case, the vector of repeated measurements for each subject is summarized by a vector of a relatively small number of estimated subject-specific regression coefficients. And in the second stage, multivariate regression techniques are used to relate these estimates to known covariates, classifications and baseline characteristics.

The well known statistical model for multilevel analysis is the hierarchical linear regression model which is essentially an extension of the familiar multiple linear regression model by including the nested random coefficients. This model is known in literature under a variety of names such as the hierarchical linear model (Bryk and Raudenbush, 1992), the variance-component model (Levin, 1999), and the random-coefficient model (Longford, 1993). The hierarchical linear model is also very useful for analysing repeated measures or longitudinal data. Since the appearance of the path-breaking paper by Laird and Ware (1982), the hierarchical linear model has been the main type of application of the hierarchical linear model in the biological and medical sciences (Snijders and Bosker, 1999). Its main advantage is that it effectively elaborates the modelling of the random variability. However, of importance is its flexibility to deal with unbalanced data structures or longitudinal data where measurements are made at different sets of time points. This is mainly possible because the model does not assume equal numbers of identical occasions for all subjects. Since the hierarchical linear regression model contains both random and fixed effects, it falls under the mixed

model family.

3.5.1 The mixed model

Models where some of the factors in the model structure are fixed and some factors are random effects are known as mixed models. The general linear mixed model can be represented in matrix form by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

in which $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants (the fixed effects), \mathbf{X} is an $n \times p$ design matrix of fixed numbers associated with $\boldsymbol{\beta}$, $\mathbf{Z} = [\mathbf{Z}_1 | \mathbf{Z}_2 | \dots | \mathbf{Z}_r]$ where \mathbf{Z}_i is an $n \times q_i$ design matrix for the random effect factor i , $\mathbf{u}' = [\mathbf{u}'_1 | \mathbf{u}'_2 | \dots | \mathbf{u}'_r]$, where \mathbf{u}_i is a $q_i \times 1$ vector of independent random variables with a $N(0, \sigma_i^2)$ distribution, $i = 1, 2, \dots, r$, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of error terms with a $N(0, \sigma_e^2)$ distribution and \mathbf{u}_i and $\boldsymbol{\varepsilon}$ are mutually independent. One may also write $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$, where \mathbf{G} is a block diagonal with the i -th block $\sigma_i^2 \mathbf{I}_{q_i}$, so that \mathbf{y} has a multivariate normal distribution with $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $Var(\mathbf{y}) = \mathbf{V}$, where $\mathbf{V} = \sigma_e^2 \mathbf{I}_n + \mathbf{Z}\mathbf{G}\mathbf{Z}' = \sigma_e^2 \mathbf{I}_n + \sum_{i=1}^r \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i'$ (Zewotir and Galpin, 2004).

In practice, the variance components σ_e^2 and σ_i^2 need to be estimated. This can be done by using either the analysis of variance (ANOVA), maximum likelihood (ML) or the restricted/residual maximum likelihood (REML) methods. For balanced data (where the number of observations in the subclasses of the data are the same), it is common practice to estimate the variance components using the ANOVA method. This is done by equating the sum of squares (or mean squares) in an ANOVA table to their expected values. However, this method has a limitation in that it is applicable only to balanced data. Another disadvantage is that it sometimes gives negative variance estimates. For unbalanced data however, Henderson (1953) describes three ways of

using the general ANOVA method for estimating the variance components. These are the Henderson's methods I, II and III.

Henderson's method I also uses the sum of squares but is adjusted to handle unequal cluster sizes. Computed are the sum of squares as in the standard ANOVA of corresponding orthogonal data. The sum of squares are then equated to their expectations, obtained under the assumption of the model, to solve for the unknown variances. Despite this method being the simplest, it can only be used if it is assumed that, except for the general/overall mean, all the elements of the model are uncorrelated variables with means zero and variances σ_e^2 and σ_i^2 . This is because it leads to biased estimates if certain elements of the model are fixed or if some of them are correlated. Its for this reason that this method cannot handle mixed models.

Henderson's method II tries to separate the random part from the general model so as to enable the application of the sum of squares method to the isolated random model. This is done by obtaining the least squared estimates of the fixed effects, correcting the data according to their fixed effects, using the corrected data in place of the original data, and then proceeding on as in Method I. In this method, the bias in estimating the variance components due to the assumption that fixed elements of the model are random variables is eliminated. At the same time the simplicity of Method I is retained. Although this method works well with mixed models, it cannot be used for models with interactions between the fixed and random factors.

Henderson's method III, on the other hand, uses the sum of squares that arise in fitting an over-parameterized model and sub-models. In this method, the mean squares are computed by a conventional least squares analysis of non-orthogonal data (method of fitting constants, weighted squares of means for example). These mean squares are then equated to their expectations to solve for the unknown variances. This method yields unbiased estimates but the computations required may be excessive. This is

because it suffers the problem of many sum of squares.

When it is computationally feasible, Method III is the most satisfactory of the three methods of estimating variance components. This is mainly because it gets around the difficulty of fixed elements in the model. In addition, it yields unbiased estimates even though certain elements in the model are correlated. Unfortunately, Method III is not likely to be computationally feasible in the non-orthogonal case unless the number of different classes is small or unless the design incorporates planned non-orthogonality such that consequently the mean squares of the ANOVA can be computed without solving least square equations. However, these methods also have the problem of sometimes giving negative variance estimates. The relative sizes of the sampling variances of estimates obtained by these three methods are also unknown.

The ML method maximizes the likelihood function using the Newton Raphson and/or Fisher's scoring algorithms to obtain the estimates. On the other hand, the REML method minimizes the likelihood function of a certain number of linearly independent error contrasts using the Newton Raphson and/or Fisher's scoring algorithms to obtain the estimates of the variance components.

According to Zewotir and Galpin (2004), working in terms of $\gamma_i = \sigma_i^2/\sigma_e^2$ eases the computational procedures associated with the model. Therefore, $Var(y)$ can be written as $Var(y) = \sigma_e^2 H$ where $H = I_n + ZDZ' = I_n + \sum_{i=1}^r \gamma_i Z_i Z_i'$ and $D = G/\sigma_e^2$ (also see Wolfinger, Tobias and Sall, 1994; Littell et al., 1996). The best linear unbiased estimate (BLUE) of β is $\hat{\beta} = (X'H^{-1}X)^{-1}X'H^{-1}y$ and the best linear unbiased predictor (BLUP) of u is $\hat{u} = DZRy$ where $R = H^{-1} - H^{-1}X(X'H^{-1}X)^{-1}X'H^{-1}$. The ML and REML estimates of σ_e^2 are $\hat{\sigma}_e^2 = y'Ry/n$ and $\hat{\sigma}_e^2 = y'Ry/(n - p)$ respectively. If the variance components γ_i are unknown, which is usually the case, the ML or REML estimated are simply substituted into the expressions. In this case the BLUE and BLUP acronyms no longer apply, but a qualifying empirical is often added to indicate such

approximation.

The relative advantage of the ML over the REML method is that it provides estimates of the fixed effects, while REML does not (Searl, Casella and McCulloch, 1992). However, REML estimators takes into account the degrees of freedom involved in estimating the fixed effects, while the ML estimators do not. For balanced data, ANOVA and REML estimates are the same whereas for unbalanced data, ML and REML estimates are slightly more efficient than the ANOVA (Henderson's methods) estimates (Snijders and Bosker, 1999).

3.5.2 The hierarchical linear model

The hierarchical linear model is the best model for multilevel data because it accounts for the within-subject as well the between-subject variations in the data. The "subjects" refer to the units at the higher level of the nesting hierarchy. Using the description of the hierarchical linear model by Snijders and Bosker (1999): for $i = 1, \dots, n$ and $j = 1, \dots, m_i$, let y_{ij} and x_{ij} be the respective measurements of the dependent and the independent variables made on the j^{th} unit in the i^{th} subject. Further, let z_i ($i = 1, \dots, n$) be the measurement of the independent variable made on the i^{th} subject.

The simplest model for y_{ij} is one without the random effect, which is the characteristic of the classical multiple regression model. This classical multiple regression model states that the dependent variable can be expressed as a linear combination of the explanatory variables and the random residual as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_i + R_{ij}, \quad (3.2)$$

where the β 's are the regression coefficients/parameters. Specifically, β_0 is the intercept (the value obtained when x_{ij} and z_i are zero); β_1 is the coefficient for the unit within subject variable X ; β_2 is the coefficient for the subject variable Z ; and variable R_{ij} is

the error term.

Model (3.2) can be extended to include the cross-level interaction effect resulting in the regression model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_i + \beta_3 z_i x_{ij} + R_{ij}. \quad (3.3)$$

The model assumes that the whole data structure can be fully explained by the subject variable Z and the unit within subject variable X . However, the nesting structure is not completely described by this model. Additional effects of the nesting structure can be incorporated by letting the regression coefficients vary from subject to subject. This results in subject-dependent coefficients β_{0i} and β_{1i} yielding the model

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + R_{ij}. \quad (3.4)$$

These subject-dependent coefficients can be modelled as follows:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} z_i + U_{0i}, \text{ and} \quad (3.5)$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} z_i + U_{1i} \quad (3.6)$$

where γ_{00} and γ_{10} are the average intercepts, γ_{01} and γ_{11} are the regression coefficients, and U_{0i} and U_{1i} are the subject-dependent deviations. If β_{0i} and β_{1i} are found to be constant, then the nesting structure has no effect and one returns to the model (3.3) by substituting the subject-dependent coefficients (3.5) and (3.6) in (3.4) without the random effects U_{0i} and U_{1i} . In this case, the OLS regression offers a good approach for data analysis. If only β_{0i} is found to be random and β_{1i} is found to be constant, then the resulting model is known as the 'random intercept model'. This is the simplest case of the hierarchical model which after substituting (3.5) and (3.6) in (3.4), without the random effect U_{1i} , is given by

$$\begin{aligned} y_{ij} &= (\gamma_{00} + \gamma_{01} z_i + U_{0i}) + (\gamma_{10} + \gamma_{11} z_i) x_{ij} + R_{ij} \\ &= \gamma_{00} + \gamma_{01} z_i + \gamma_{10} x_{ij} + \gamma_{11} z_i x_{ij} + U_{0i} + R_{ij}. \end{aligned}$$

However, a more general case of the hierarchical linear model is where the slopes are also random (represented in Model (3.4)). Substituting (3.5) and (3.6) in model (3.4) leads to the model

$$\begin{aligned} y_{ij} &= (\gamma_{00} + \gamma_{01}z_i + U_{0i}) + (\gamma_{10} + \gamma_{11}z_i + U_{1i})x_{ij} + R_{ij} \\ &= \gamma_{00} + \gamma_{01}z_i + \gamma_{10}x_{ij} + \gamma_{11}z_ix_{ij} + U_{0i} + U_{1i}x_{ij} + R_{ij}. \end{aligned}$$

This model can be extended to include more subject variables and more unit within subject variables. Suppose that there are p unit level explanatory variables X_1, \dots, X_p and q subject level explanatory variables Z_1, \dots, Z_q , then the regression model 3.4 becomes

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{1ij} \dots + \beta_{pi}x_{pij} + R_{ij}. \quad (3.7)$$

The subject-dependent coefficient β_{hi} ($h = 1, \dots, p$), which can be known as the q -variable regression model for the i -th subject, is given by

$$\beta_{hi} = \gamma_{h0} + \gamma_{h1}z_{1i} + \dots + \gamma_{hq}z_{qi} + U_{hi}. \quad (3.8)$$

Substituting (3.8) in (3.7) and rearranging the terms in the model yields

$$y_{ij} = \gamma_{00} + \sum_{h=1}^p \gamma_{h0}x_{hij} + \sum_{k=1}^q \gamma_{0k}z_{ki} + \sum_{k=1}^q \sum_{h=1}^p \gamma_{hk}z_{ki}x_{hij} + U_{0i} + \sum_{h=1}^p U_{hi}x_{hij} + R_{ij}.$$

The between subject variation is now characterized by $p+1$ random coefficients, $U_{0i}, U_{1i}, \dots, U_{pi}$. These random coefficients are independent between subjects, but may be correlated within subjects. It is assumed that the vector $(U_{0i} \dots U_{pi})$ is independent of the unit level error terms R_{ij} and that all error terms have mean 0, given the values of all the explanatory variables. It is also assumed that the error term R_{ij} has a normal distribution with constant variance σ^2 and $(U_{0i} \dots U_{pi})$ has a multivariate normal distribution with a constant covariance matrix with the variances and covariances of

the subject random effects denoted as:

$$\begin{aligned} \text{Var}(U_{hi}) &= \tau_{hh} = \tau_h^2 \quad (h = 1, \dots, p), \\ \text{Cov}(U_{hi}, U_{ki}) &= \tau_{hk} \quad (h, k = 1, \dots, p), \\ \text{Var}(U_{0i}) &= \tau_{00} = \tau_0^2, \text{ and} \\ \text{Cov}(U_{0i}, U_{hi}) &= \tau_{0h} \quad (h = 1, \dots, p). \end{aligned}$$

Snijders and Bosker(1999), and Efron and Morris (1975) highlight a method known as the Empirical Bayes (EB) estimation which produces posterior means used to predict U_{0i} (also known as latent variables). The basic idea of this method is that U_{0i} is estimated by combining two kinds of information: the data from subject i ; and the model assumption that the unobserved U_{0i} is a random variable, therefore has a normal distribution with mean 0 and variance τ_0^2 .

The hierarchical linear model also provides room for statistical testing, also known as model fitting, with the sole aim of obtaining a parsimonious model that fits reasonably well with the data. To get a parsimonious model, all the effects are tested and only the significant ones (at a predetermined level of significance) are included in the model.

Chapter 4

The longitudinal mixed model

Verbeke and Molenberghs (2000) describe the longitudinal mixed model by combining two stages of analysis into one single statistical model as follows. Let the random variable Y_{ij} denote the response of interest of the i -th subject, measured at time j , $i = 1, \dots, N$; $j = 1, \dots, n_i$ and let \mathbf{y}_i be the n_i -dimensional vector of all repeated measurements on the i -th subject that is, $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})'$. Then the first stage of the two-stage approach assumes that \mathbf{y}_i satisfies the linear regression model

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \quad (4.1)$$

where \mathbf{Z}_i is an $(n_i \times q)$ matrix of known covariates modelling how the response evolves over time for the i -th subject, while $\boldsymbol{\beta}_i$ is a q -dimensional vector of unknown subject-specific regression coefficients and $\boldsymbol{\varepsilon}_i$ is a vector of the error term components, ε_{ij} , $j=1, \dots, n_i$. It is usually assumed that all $\boldsymbol{\varepsilon}_i$ are independent and normally distributed with mean vector zero and covariance matrix $\boldsymbol{\Sigma}_i$, where $\boldsymbol{\Sigma}_i$ is the n_i -dimensional matrix. In the second stage, a multivariate regression model, of the form

$$\boldsymbol{\beta}_i = \mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i, \quad (4.2)$$

is used to explain the observed variability between the subjects with respect to their subject-specific regression coefficients β_i with K_i as a $(q \times p)$ matrix of known covariates and β as the p -dimensional vector of unknown regression parameters. Lastly, the b_i 's are assumed to follow a q -dimensional normal distribution with mean vector zero and general covariance matrix D .

The regression parameters in the second stage model (4.2), which are of primary interest, can be estimated by sequentially fitting the stage one and two models. First, all β_i are estimated by fitting model (4.1) to the observed data vector y_i of each subject separately, yielding estimates $\hat{\beta}_i$. Then model (4.2) is fitted to the estimates $\hat{\beta}_i$, providing inferences for β .

As for any analysis of summary statistics, the two-stage analysis obviously suffers from problems. Firstly, information is lost in summarizing the vector y_i of the observed measurements on the i -th subject by $\hat{\beta}_i$. Secondly, the random variability is introduced when β_i in model (4.2) is replaced by its estimate $\hat{\beta}_i$. Moreover, the covariance matrix of $\hat{\beta}_i$ depends strongly on the number of measurements made on the i -th subject as well as on the time points at which these measurements were taken. This is not taken into account at the second stage of analysis, but can be resolved by combining the two stages of analysis into one model commonly known as the linear mixed (-effects) model. The combination is done by replacing β_i in model (4.1) by its expression in model (4.2), yielding:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \quad (4.3)$$

$$b_i \sim N(0, D), \quad (4.4)$$

$$\varepsilon_i \sim N(0, \Sigma_i), \quad (4.5)$$

where $b_1, \dots, b_N, \varepsilon_1, \dots, \varepsilon_N$ are independent, with D as a general $(q \times q)$ covariance matrix with (i, j) element $d_{ij} = d_{ji}$ and Σ_i is an $(n_i \times n_i)$ covariance matrix respectively.

Matrices X_i (where $X_i = Z_i K_i$) and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates respectively. The p -dimensional vector β represents the fixed effects while the q -dimensional vector b_i are the subject-specific random effects. Although the random effects and the error term components are independent of each other, the subject-effects and the random effects are usually not independent but are correlated (Hallahar, 2004). The response vector y_i is normally distributed with mean vector $X_i \beta$ and covariance matrix $V_i = Z_i D Z_i' + \Sigma_i$.

4.1 The fixed-effects/mean structure

In order to choose an appropriate fixed-effects structure for the longitudinal mixed model, one needs to examine how the profile for the relevant sub-populations of the subjects evolve over time. This enables the visualization of the data pattern over time. Individual and mean profiles of the subjects over time are commonly used to visualize these data patterns over time. Individual profiles are the responses of each subject plotted against time, while in the mean profiles are the mean responses of the subjects are plotted against time.

However, it is difficult to determine the most appropriate functional form of the dependence of the response variable on time by merely looking at the profile plots. To solve this problem, a scatter-plot smoother known as the loess smoother can be used. This smoother produces non-parametric curves that show the functional dependence without imposing parametric assumptions about the dependence. The loess smoother can be implemented by using the SAS PROC LOESS procedure. As described in SAS Institute Inc. (2000) and by Cohen (1999), PROC LOESS implements a non-parametric method for estimating local regression surfaces. This method is commonly referred to as 'loess', which is the abbreviation for local regression. Assume that for i

= 1 to n , the i th measurement y_i of the response y and the corresponding measurement x_i of the vector x of p predictors are related by

$$y_i = g(x_i) + \epsilon_i,$$

where g is the regression function and ϵ_i is a random error. The idea of local regression is that the regression function $g(x)$ can be locally approximated by the value of a function in some specified parametric class near $x = x_0$. Such a local approximation is obtained by fitting a regression surface to the data points within a chosen neighbourhood of the point x_0 . In the loess method, weighted least squares are used to fit linear or quadratic functions of the predictors at the centers of neighbourhoods. The radius of each neighbourhood is chosen so that the neighbourhood contains a specified percentage of the data points. The fraction of the data, called the smoothing parameter, in each local neighbourhood controls the smoothness of the estimated surface. Data points in a given local neighbourhood are weighted by a smooth decreasing function of their distance from the center of the neighbourhood.

The loess fit, which is usually applied to a random sample of subjects (in large sample surveys), depends strongly on a smoothing parameter. The most appropriate smoothing parameter is one which minimizes a criterion that incorporates both the tightness of the fit and the model complexity. To determine such a smoothing parameter, the Akaike information criteria (AIC) is used. The smoothing parameter that provides the smallest AIC value is selected as the most appropriate. The values of AIC for all the potential smoothing parameters can be obtained from the "Fit summary" table of PROC LOESS provided that one of the options ALL, CLM, DFMETHOD=EXACT, STD, and T are specified in the model statement.

The quality of the estimates in a model, as measured by their variances, can be adversely affected if the independent variables are closely related (that is highly cor-

related) to each other (Sen and Srivastava, 1990). The more correlated the variables are, the more difficult it is to determine which of the related variables is producing the effect on the response variable. This problem is referred to as multicollinearity. Large absolute correlation coefficients between the independent variables indicate the presence of multicollinearity.

Another method for detecting multicollinearity is to assess the degree to which each independent variable is related to all other independent variables (Sen and Srivastava, 1990). If there are p independent variables, then this is done by examining the R_j^2 ($j = 1, \dots, p$), which is the sample coefficient of determination (R^2), between the variable x_j and all the other independent variables. The tolerance TOL_j is defined as

$$TOL_j = 1 - R_j^2.$$

If x_j is not closely related to the other variables then tolerance TOL_j will be close to 1. The variance inflation factor VIF_j is given by

$$VIF_j = TOL_j^{-1}.$$

Clearly, a value of VIF_j close to 1 indicates no linear relationship among the independent variables, while larger values indicate presence of multicollinearity. VIF s larger than 10 imply serious problems with multicollinearity (Montgomery, Peck and Vining, 2001).

In a case of multicollinearity, the principal component analysis (PCA) can be applied on the data. In addition, a rotation can be applied to the PCA process. Its goal is to minimize the complexity of the component by making the large loadings larger and the small loadings smaller within each component (Wuensch, 2004). The most commonly used rotation is the VARIMAX rotation. However there are other rotation methods such as the QUARTIMAX and the EQUAMAX rotation. The QUARTIMAX rotation makes large loadings larger and small loadings smaller within each variable while the

EQUAMAX rotation is a compromise that attempts to simplify both components and variables. These are all orthogonal rotations, that is, the axes remain perpendicular, so the components are not correlated with one another.

4.2 The covariance structure

Most models assume that ε_i has a constant variance and can be decomposed as $\varepsilon_i = \varepsilon_{(1)i} + \varepsilon_{(2)i}$ in which $\varepsilon_{(1)i}$ is the component of the measurement error and $\varepsilon_{(2)i}$ is a component of serial correlation (Diggle *et al.*, 1994). This suggests that at least part of a subject's observed profile is a response to the time-varying stochastic processes operating within the subject. This type of random variation results in a correlation between serial measurements which are usually a decreasing function of the time separation between the repeated measurements. The component $\varepsilon_{(1)i}$ is an extra component of measurement error reflecting variation added by the measurement process itself. This yields three stochastic components which are the random effects, serial correlation effect and the measurement error, resulting in the linear mixed model

$$y_i = X_i\beta + Z_i b_i + \varepsilon_{(1)i} + \varepsilon_{(2)i}$$

where:

$$b_i \sim N(0, D),$$

$$\varepsilon_{(1)i} \sim N(0, \sigma^2 \mathbf{I}_{n_i}),$$

$$\varepsilon_{(2)i} \sim N(0, \tau^2 \mathbf{H}_i),$$

with $b_1, \dots, b_N, \varepsilon_{(1)1}, \dots, \varepsilon_{(1)N}, \varepsilon_{(2)1}, \dots, \varepsilon_{(2)N}$ independent and the correlation matrix \mathbf{H}_i assumed to be a $(n_i \times n_i)$ structure. However, in many applications, the effect of serial correlation is very often dominated by the combination of random effects and measurement error. This has led to fitted models in practice not including the serial

correlation. To check whether there is a need for serial correlation effect in the model, scatter plots of the residuals are constructed. Scatter plots of a circular shape indicate that serial correlation is not very strong (Fanta, 2003).

4.3 Modelling the covariance structure

In modelling the covariance structure, the SAS PROC MIXED procedure provides a rich assortment of covariance structures from which to select. However, three of the most commonly used structures in longitudinal mixed models are the compound symmetric (CS), autoregressive order one (AR(1)) and the unstructured (UN). To define these structures in terms of the longitudinal mixed model shown in (4.3), let R define the block diagonal of the covariance matrix Σ_i with each block corresponding to a subject, then the variance-covariance matrix of the observations is given by $\text{var}(y) = Z'GZ + R$.

In the case of the CS structure, $\Sigma_i = (J\rho + I(1-\rho))\sigma^2$ where I_{n_i} is an $n_i \times n_i$ identity matrix, J_{n_i} is an $n_i \times n_i$ matrix of ones (unit matrix) and ρ is the correlation between observations on the the i -th subject.

In the case the of AR(1) structure, the correlation between observations, which are say w time periods apart, is assumed to be ρ^w , where ρ is the correlation between the adjacent observations on the i -th subject . Therefore Σ_i is given by:

$$\Sigma_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \dots & \rho^{n_i-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{n_i-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho^{n_i-1} & \dots & \dots & \dots & \rho & 1 \end{pmatrix}.$$

The correlations between the observations are dependent on the time periods between the observations.

For the UN structure, there are no mathematical constraints imposed on the elements of the covariance matrix. Thus unstructured means no structure. The problem now is how to decide which of the three covariance structures to assume in the model. This decision process can be assisted by using three model-fit criteria which are: the Akaike's Information Criterion (AIC), the finite-sample corrected version of AIC (AICC) and Schwarz' Bayesian Information Criterion (BIC). These are essentially log likelihood values penalized for the number of parameters estimated, hence the criteria reported here should only be used to compare models with the same mean structure but with different covariance structures. The BIC however imposes a heavier penalty than AIC. The covariance structure with the largest values of the criteria is considered most desirable. These values can be obtained from the "Fit Statistics" table after applying the SAS PROC MIXED procedure. The interest in the covariance structure is not for its own right but for obtaining a good model for the covariance structure so that computations and inferences about the fixed effects are valid.

4.4 Model reduction

Over-parameterization of the model structure leads to inefficient estimation and potentially poor assessment of standard errors for the estimates. This results in the need for model reduction of both the fixed and random parameters and this is commonly done by model comparisons.

4.4.1 Tests for the significance of the fixed effects

In the model reduction of fixed effects, a full model of all possible effects is compared with reduced models in order to obtain the most appropriate parsimonious model. To compare the models, the likelihood ratio (LR) test is used. This is a classical sta-

tistical test for the comparison of nested models with different mean structures. The LR test statistic is defined as:

$$-2 \ln \lambda_N = -2 \ln \left[\frac{L_{ML}(\hat{\theta}_{ML,0})}{L_{ML}(\hat{\theta}_{ML})} \right], \quad (4.6)$$

where $\hat{\theta}_{ML,0}$ and $\hat{\theta}_{ML}$ are the respective maximum likelihood estimates which maximize the ML likelihood functions of the full and reduced models. The LR test statistic has a chi-square distribution with the degrees of freedom equal to the difference between the number of parameters in the two models. The above result is not valid if the models are fitted using the REML method (Snijders and Bosker, 1999). The -2 times the log likelihoods for the full and the reduced models can be obtained from the "Fit Statistics" table after running PROC MIXED for each of the two models. Then the LR test statistic is calculated using equation 4.6.

The reduced model can be reduced even further by applying the backward elimination method. In this method, all the potential fixed effect variables are introduced in the model and the one with the largest p -value (or in other words the smallest F -value) is identified (Bowerman, O'Connell, and Dickey, 1986 pg 342; and Neter, Wasserman, and Kutner, 1990 pg 458). If this p -value is larger than α (where α is the level of significance), then this variable is removed /dropped from the model. The remaining variables are tested again and the variable with a p -value greater than α is removed from the model. The process continues until all the variables remaining in the model have p -values less than α .

4.4.2 Tests for the significance of the random effects

The LR test, however, does not exist for cases where the null hypothesis specifies that the parameter lies on the boundary of the parameter space, which is typically for variance component models for which the standard asymptotic theory does not apply

(Self and Liang, 1987). Using results of Self and Liang (1987) on nonstandard testing situations, Stram and Lee (1994, 1995) were able to show that the asymptotic null distribution for the LR test statistic for testing the hypothesis of the need for random effects is often a mixture of chi-squared distributions rather than the classical single chi-squared distribution. This is derived under the assumption of conditional independence which states that all residual covariances Σ_i are of the form $\sigma^2 \mathbf{I}_{n_i}$. Stram and Lee (1994, 1995) discuss the following specific LR tests.

Case 1: No random effects versus one random effect: For testing $H_0: D = 0$ versus $H_A: D = d_{11}$, where d_{11} is a non-negative scalar that represents the variance component of the random effect, then the asymptotic null distribution of $-2\ln \lambda_N$ is a mixture of χ_1^2 and χ_0^2 with equal weights 0.5. The χ_0^2 distribution is the distribution which gives probability mass 1 to the value 0.

Case 2: one versus two random effects: In this case, one wishes to test

$$H_0 : D = \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

for strictly positive d_{11} , versus H_A that D is a (2×2) positive semi-definite matrix. The asymptotic null distribution of $-2\ln \lambda_N$ is a mixture with equal weights 0.5 for χ_2^2 and χ_1^2 .

Case 3: q versus $q + 1$ random effects: In this case, one wishes to test

$$H_0 : D = \begin{pmatrix} D_{11} & \mathbf{0} \\ \mathbf{0}' & 0 \end{pmatrix},$$

in which D_{11} is a $(q \times q)$ positive definite matrix, versus H_A that D is a general $((q + 1) \times (q + 1))$ positive semi-definite matrix. The large-sample behaviour of the null

distribution of $-2\ln \lambda_N$ is a mixture of χ_{q+1}^2 and χ_q^2 , again with equal weights 0.5.

Case 4: q versus $q + k$: In this case, one wishes to test the H_0 in case 3 versus

$$H_A : \mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}'_{12} & \mathbf{D}_{22} \end{pmatrix},$$

in which \mathbf{D} is a general $((q + k) \times (q + 1))$ positive semi-definite matrix. The null distribution of $-2\ln \lambda_N$ is a mixture of χ^2 random variables formed by the lengths of projections of multivariate normal random variables upon curved as well as flat surfaces.

Note that, for all the cases above, if the classical null distribution is used, then all p -values would be overestimated. Therefore the null hypothesis would be accepted too often, resulting in incorrectly simplifying the covariance structure of the model, thus invalidating inferences (Altham, 1984). The correction for the boundary parameter values under the null hypotheses therefore reduces the p -values in order to protect against the use of an oversimplified or a too parsimonious covariance structure.

Although the results in Stram and Lee (1994, 1995) were derived for the case of ML estimation, the same results apply for REML estimation (Morrell, 1998). In fact, the REML test statistic performs slightly better than the ML test statistic in the sense that, on average, the rejection proportions are closer to the nominal level for the REML test statistic than for the ML test statistic (Verbeke and Molenberghs, 2000). For the longitudinal mixed model, testing is done by deleting one random effect at a time from the model starting with the highest-order effect and testing for significance of whether the deleted random effect is needed in the model.

4.5 Estimation of the fixed and random effects

Let α denote the vector of all the variance and covariance parameters (commonly known as variance components) in $V_i = Z_i D Z_i' + \Sigma_i$. And let $\theta = (\beta', \alpha')$ be the vector of all parameters in the marginal model for y_i . Then the classical approach to inference is based on estimators obtained from maximizing the marginal likelihood function

$$L_{ML}(\theta) = \prod_{i=1}^N \left\{ (2\pi)^{-n_i/2} |V_i(\alpha)|^{-1/2} \times \exp \left(-\frac{1}{2} (y_i - X_i \beta)' V_i^{-1}(\alpha) (y_i - X_i \beta) \right) \right\} \quad (4.7)$$

with respect to θ . If α is known, then the maximum likelihood estimator (MLE) of β , obtained by maximizing (4.7) and conditional on α , is given by

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^N X_i' W_i X_i \right)^{-1} \sum_{i=1}^N X_i' W_i y_i, \quad (4.8)$$

where $W_i = V_i^{-1}$ (Laird and Ware, 1982). However if α is not known but an estimate $\hat{\alpha}$ is available, then W_i can be replaced by its estimate $\hat{W}_i = \hat{V}_i^{-1} = V_i^{-1}(\hat{\alpha})$. Two frequently used methods for estimating α are the MLE and the REML. The MLE of α is obtained by maximizing (4.7) with respect to α after the β has been replaced by (4.8).

In practice, the linear mixed models often contain many fixed effects. In such cases, it is important to estimate the variance components explicitly, taking into account the loss of the degrees of freedom involved in estimating the fixed effects. The REML method puts this into practice by applying the error contrasts as follows. First the N subject-specific regression models are combined into one model given by

$$y = X\beta + Zb + \varepsilon, \quad (4.9)$$

where the vectors y , b and ε and the matrix X are obtained by stacking the vectors y_i , b_i and ε_i and matrices X_i on each other. The matrix Z is the block-diagonal

matrix with blocks Z_i on the main diagonal and zero elsewhere. The dimension of \mathbf{y} is $n = \sum_{i=1}^N n_i$. The distribution of \mathbf{y} is multivariate normal with mean $\mathbf{X}\beta$ and covariance matrix $V(\alpha)$. The covariance matrix $V(\alpha)$ is also a block-diagonal matrix with blocks V_i on the main diagonal and zeros elsewhere. The REML estimator for the variance components α is obtained from maximizing the likelihood function of a set of error contrasts $\mathbf{U} = \mathbf{A}'\mathbf{y}$ where \mathbf{A} is any $(n \times (n - p))$ full-rank matrix with its columns orthogonal to the columns of the matrix \mathbf{X} . The vector \mathbf{U} has a normal multivariate distribution with mean vector zero and covariance matrix $\mathbf{A}'V(\alpha)\mathbf{A}$ which is independent of β . The likelihood function of the error contrasts is given by

$$\begin{aligned} L(\alpha) &= (2\pi)^{-(n-p)/2} \left| \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right|^{1/2} \\ &\times \left| \sum_{i=1}^N \mathbf{X}'_i \mathbf{V}_i^{-1} \mathbf{X}_i \right|^{-1/2} \prod_{i=1}^N |\mathbf{V}_i|^{-1/2} \\ &\times \exp \left\{ -\frac{1}{2} \sum_{i=1}^N (\mathbf{y}_i - \mathbf{X}_i \hat{\beta})' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\beta}) \right\}, \end{aligned}$$

where $\hat{\beta}$ is given by (4.8) (Harville, 1974). This shows that the REML estimator $\hat{\alpha}$ does not depend on the error contrasts (that is the choice of \mathbf{A}).

This likelihood function also equals

$$L(\alpha) = C \underbrace{\left| \sum_{i=1}^N \mathbf{X}'_i \mathbf{W}_i(\alpha) \mathbf{X}_i \right|^{-1/2}}_I L_{ML}(\hat{\beta}(\alpha), \alpha), \quad (4.10)$$

where C is a constant with respect to both α and β , and $L_{ML}(\beta, \alpha) = L_{ML}(\theta)$ is the ML likelihood function in (4.7). Therefore, the REML estimators for α and β can also be found by maximizing the so-called REML likelihood function

$$L_{REML}(\theta) = \left| \sum_{i=1}^N \mathbf{X}'_i \mathbf{W}_i(\alpha) \mathbf{X}_i \right|^{-1/2} L_{ML}(\theta)$$

with respect to all parameters (α and β) simultaneously.

The main justification of the REML approach is that in the absence of information on β , no information about α is lost when inference is based on the vector \mathbf{U} rather than

on \mathbf{y} (Patterson and Thompson, 1971). From a Bayesian point of view, using only error contrasts to make inferences about α is equivalent to ignoring any prior information about β , and using all the data to make inferences about α (Harville, 1974). However, both ML estimation and REML estimation have the same merits of being based on the likelihood principle which leads to useful properties such as consistency, asymptotic normality and efficiency.

In literature, several methods for the actual calculation of the ML or REML estimates have been described. Dempster, Laird and Rubin (1977) for example, introduced the Expectation-Maximization (EM) algorithm for the calculation of the MLEs based on incomplete data and illustrated how it can be used for the estimation of variance components in mixed-models. Not only can the EM algorithm be used to calculate the MLEs, but it can also be used to calculate the REML estimates through the empirical Bayesian approach. However, slow convergence of the algorithm to the estimate of variance components is experienced especially when the maximum likelihood estimates are on or near the boundary of the parameter space (Laird and Ware, 1982). Therefore to circumvent these convergence problems, researchers use the Newton-Raphson (and/or the Fisher's scoring)-based procedures to estimate all parameters in the model.

It is also useful to calculate estimates for the random effects \mathbf{b}_i since they reflect how much the subject-specific profiles deviate from the overall average profile. Since the random effects in the linear mixed effects model are assumed to be random variables, it is most natural to estimate them using the Bayesian techniques. Very often \mathbf{b}_i is estimated by the mean of a posterior distribution called the posterior mean of \mathbf{b}_i . This

estimate is given by:

$$\begin{aligned}\widehat{\mathbf{b}}_i(\theta) &= E[\mathbf{b}_i | \mathbf{Y}_i = \mathbf{y}_i] \\ &= \int \mathbf{b}_i f(\mathbf{b}_i | \mathbf{y}_i) d\mathbf{b}_i \\ &= \mathbf{DZ}_i \mathbf{W}_i(\alpha)(\mathbf{y}_i - \mathbf{X}_i \beta),\end{aligned}$$

where \mathbf{Y}_i is the random vector of responses on subject i and \mathbf{y}_i are the observed values of \mathbf{Y}_i . The covariance matrix of the corresponding estimator is given by

$$\text{Var}(\widehat{\mathbf{b}}_i(\theta)) = \mathbf{DZ}'_i \left\{ \mathbf{W}_i - \mathbf{W}_i \mathbf{X}_i \left(\sum_{i=1}^N \mathbf{X}'_i \mathbf{W}_i \mathbf{X}_i \right)^{-1} \mathbf{X}'_i \mathbf{W}_i \right\} \mathbf{Z}_i \mathbf{D},$$

which underestimates the variability of $\widehat{\mathbf{b}}_i(\theta) - \mathbf{b}_i$ since it ignores the variation of \mathbf{b}_i .

Therefore the inference of \mathbf{b}_i is usually based on

$$\text{Var}(\widehat{\mathbf{b}}_i(\theta) - \mathbf{b}_i) = \mathbf{D} - \text{Var}(\widehat{\mathbf{b}}_i(\theta))$$

as an estimator for the variation of $\widehat{\mathbf{b}}_i(\theta) - \mathbf{b}_i$ (Liard and Ware, 1982). The unknown parameters β and α are replaced by their ML or REML estimates. The resulting estimates for the random effects are called the Empirical Bayes (EB) estimates which can be denoted as $\widehat{\mathbf{b}}_i$. However, the true variability in the estimate $\widehat{\mathbf{b}}_i$ is underestimated since the variability introduced by replacing the unknown parameter θ by its estimate is not taken into account.

Estimates of the \mathbf{b}_i can also be obtained from solving a system of linear equations (Henderson *et al.*, 1959). Let the linear mixed model be as given by (4.9). If the variances \mathbf{D} and Σ are block-diagonal matrices with the blocks in \mathbf{D} and in Σ_i on the main diagonal and zero elsewhere, then the estimates of the fixed effects β and all the random effects represented by vector \mathbf{b} can be obtained from solving the mixed model normal equations:

$$\begin{pmatrix} \mathbf{X}'\Sigma^{-1}\mathbf{X} & \mathbf{X}'\Sigma^{-1}\mathbf{Z} \\ \mathbf{Z}'\Sigma^{-1}\mathbf{X} & \mathbf{Z}'\Sigma^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\Sigma^{-1}\mathbf{y} \\ \mathbf{Z}'\Sigma^{-1}\mathbf{y} \end{pmatrix}.$$

For large data sets, the computation of these equations may be expensive, making the Bayesian approach more efficient.

Suppose that interest is in the estimation of a linear combination $\mathbf{u} = \lambda'_\beta \boldsymbol{\beta} + \lambda'_b \mathbf{b}_i$ of the vector $\boldsymbol{\beta}$ of fixed effects and the vector \mathbf{b}_i of the random effects for some known vectors λ_β and λ_b of dimension p and q respectively. Then the estimator of \mathbf{u} is given by

$$\hat{\mathbf{u}}(\alpha) = \lambda'_\beta \hat{\boldsymbol{\beta}}(\alpha) + \lambda'_b \hat{\mathbf{b}}_i(\hat{\boldsymbol{\beta}}(\alpha), \alpha),$$

where $\hat{\boldsymbol{\beta}}(\alpha)$ is the ML estimator of the fixed effects and $\hat{\mathbf{b}}_i(\boldsymbol{\beta}, \alpha) = \hat{\mathbf{b}}_i(\boldsymbol{\theta})$ is the EB estimator of the random effects. The estimator $\hat{\mathbf{u}}(\alpha)$ is the best linear unbiased predictor (BLUP) of \mathbf{u} .

4.6 Test for normality

One may, however, ask oneself if the normality assumption for the random effects and the residuals is appropriate. Normality checks using the probability plots, and tests using the Shapiro-Wilk test and the Kolmogorov-Smirnov test have been used over the years. Specifically the W -statistic (in the Shapiro-Wilk test), suggested by Shapiro and Wilk (1965) has been shown to be a good omnibus test of normality (Pearson, D'Agostino and Bowman, 1977). However, using simulations and analysis of a real data, Zewotir and Galpin (2004) showed that the inferences are hardly affected by the non-normality of the random effects. The inferences however, seemed to be sensitive to the distribution of the residual term only. This restricted sensitivity led to the conclusion that it is the normality of the residual term that is important in the inferences.

The above mentioned normality tests are however based on independence of the random effects and residuals, an assumption often violated by longitudinal data. In

practice, histograms and scatter plots of components of \hat{b}_i (EB estimates) and the residuals are often used for diagnostic purposes. Specifically, the scatter plots are used to pinpoint outlying observations which arise from subjects that seem to evolve differently from the other subjects in the sample. On the other hand, the histograms of the EB estimates and the residuals can be used to check for the normality of the random effects and the residuals respectively. The histogram has shown to be more informative than the normal tests for longitudinal data (Hallahar, 2004).

4.7 Inference for fixed and random effects

In practice, the fitting of a model is rarely the ultimate goal of a statistical analysis. Instead, the primary interest is to draw inferences about the parameters in a model in order to generalize results obtained from a specific sample to the general population from which the sample was taken. Both the fixed and the random parameters are tested to check for their significance in the model. The null hypothesis that a certain model parameter is 0, that is

$$H_0 : \gamma_h = 0,$$

can be tested using a t -test. The statistical estimation procedure obtains an estimate $\hat{\gamma}_h$ with an associated standard error $S.E.(\hat{\gamma}_h)$. Their ratio is a t -value

$$T(\gamma_h) = \frac{\hat{\gamma}_h}{S.E.(\hat{\gamma}_h)}.$$

One-sided as well as two-sided tests can be carried out on the basis of this test statistic. Under the null hypothesis, $T(\gamma_h)$ has approximately a t -distribution but with the degrees of freedom (df) different from those in multiple linear regression due to the presence of the fixed and random effects.

The estimate of the fixed effect in (4.8) follows a multivariate normal distribution with mean vector β and variance-covariance matrix

$$\text{Var}(\hat{\beta}) = \left(\sum_{i=1}^N X_i' W_i X_i \right)^{-1} \left(\sum_{i=1}^N X_i' W_i \text{Var}(y_i) W_i X_i \right) \left(\sum_{i=1}^N X_i' W_i X_i \right)^{-1} \quad (4.11)$$

$$= \left(\sum_{i=1}^N X_i' W_i X_i \right)^{-1} \quad (4.12)$$

Inferences about this parameter can be done using the approximate wald test and the approximate t -tests and F -tests. The wald test, also known as the Z -test, is associated with the confidence interval of β_j , and is obtained by approximating the distribution $(\hat{\beta}_j - \beta_j) / \widehat{s.e.}(\hat{\beta}_j)$ with a standard univariate normal distribution for each parameter β_j , $j=1, \dots, p$. In general for any known full row rank matrix L , a test of the hypothesis

$$H_0: L\beta = 0 \text{ versus } H_A: L\beta \neq 0,$$

follows from the fact that the distribution of

$$(\hat{\beta} - \beta)' L' \left[L \left(\sum_{i=1}^N X_i' V_i^{-1}(\hat{\alpha}) X_i \right)^{-1} L' \right]^{-1} L(\hat{\beta} - \beta) \quad (4.13)$$

is approximately a chi-square distribution with rank (L) degrees of freedom. However, the wald test statistic is based on estimated standard errors which underestimate the true validity of $\hat{\beta}$. This is because they do not take into account the variability introduced by α (Dempster, Rubin and Tsutakawa, 1981). In practice, this downward bias is often resolved by using approximate t - and F -tests for testing the hypothesis about β . For each parameter β_j in vector β , $j=1, \dots, p$, an approximate t -test and associated confidence interval can be obtained by approximating the distribution of $(\hat{\beta}_j - \beta_j) / \widehat{s.e.}(\hat{\beta}_j)$ with a t -distribution. Testing the general linear hypothesis is thus based on an F -approximation to the distribution of

$$F = \frac{(\hat{\beta} - \beta)' L' \left[L \left(\sum_{i=1}^N X_i' V_i^{-1}(\hat{\alpha}) X_i \right)^{-1} L' \right]^{-1} L(\hat{\beta} - \beta)}{\text{rank}(L)}, \quad (4.14)$$

with the numerator degrees of freedom equal to $\text{rank}(L)$, while the denominator degrees of freedom have to be estimated from the data. The same is true for the degrees of freedom in the t -approximation. In practice, several methods are available for estimating the appropriate number of degrees of freedom needed for a specific t - or F -test. For example, the SAS PROC MIXED procedure includes five different estimation methods, one of which is based on the so-called Satterthwaite-type approximation. Kenward and Roger (1997) proposed a scaled wald statistic, based on an adjusted covariance estimate which accounts for the extra variability introduced by estimating α . They also show that its small sample distribution can be well approximated by an F -distribution with denominator degrees of freedom as obtained via a Satterthwaite-type approximation.

A sufficient condition for $\hat{\beta}$ to be unbiased is that the mean $E(y_i)$ and marginal covariance matrix are correctly specified as $X_i\beta$ and $V_i = Z_i D Z_i' + \Sigma_i$. Thus the estimate $\hat{\beta}$ is not robust with respect to model misspecification of the covariance structure. Liang and Zeger (1986) therefore propose inferential procedures based on the so-called sandwich estimator of $\text{var}(\hat{\beta})$ obtained by replacing $\text{var}(y_i)$ by $r_i r_i'$ where $r_i = y_i - X_i \hat{\beta}$. The resulting estimator is also called the robust or empirical variance estimator, and is consistent if the mean is correctly specified in the model. This suggests that as long as interest is only in the inference about $\hat{\beta}$, one may put little effort in modelling the covariance structure provided the data set is sufficiently large. However, in practice, an appropriate covariance model may be of interest since it helps in interpreting the random variation in the data. Thus, robust versions of the approximate wald test, t -test and F -test as well as the associated confidence intervals can also be obtained by replacing the covariance matrix (4.12) in (4.13) and (4.14) with the robust covariance matrix in (4.11). Note that the robust standard errors are larger, thus leading to larger confidence intervals (Verbeke and Molenberghs, 2000).

4.8 Longitudinal mixed model with missing observations

The problem of missing observations is common throughout statistical work, and is almost always present in the analysis of longitudinal or repeated measurement data. Missing data are indeed common in clinical trials, epidemiological studies, and feature very prominently in sample surveys. The most frequently encountered type of missingness in longitudinal modelling is “dropouts” which refers to the case where all observations on a subject are obtained until a certain point in time, after which all measurements are missing. Much of the treatment of missing data for longitudinal data is restricted to dropouts. Verbeke and Molenberghs (2000) highlighted the four reasons for this restriction. Firstly, the classification of the missing observation processes has a simpler interpretation with dropouts than for patterns with intermittent missing observations. Secondly, it is easier to formulate models for dropout. Thirdly much of the literature on missing observations in longitudinal data are restricted to this setting. Finally, dropouts are by far the most common modes of missingness in longitudinal studies. Missing values therefore occur as dropouts, otherwise one says that the missing values are intermittent (Diggle *et al.*, 2002).

Early work on missing values was largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design. General algorithms, such as the Expectation-Maximization (EM) (Dempster *et al.*, 1977), and data imputation and augmentation procedures (Rubin 2004, Tanner and Wong 1987), combined with powerful computing resources have largely provided solutions to this aspect of the problem. However, the difficult and important question of assessing the impact of missing data on subsequent statistical inferences still remains. Conditions can be formulated under which an analysis that proceeds as if

the missing data are missing by design (that is ignoring the missing value process) can provide valid answers to study questions.

The two common approaches to the handling of missing data are: to discard subjects with incomplete data; and simple imputation. The first approach has the advantage of simplicity, although the wide availability of more sophisticated methods of analysis minimizes the significance of this approach. It is also an inefficient use of information since some of the collected data is not used at all.

There are several forms of simple imputation. For example, a cross-sectional approach replaces a missing observation by the average of the available observations at the same time from other subjects with the same covariates and treatment. A simple longitudinal approach carries the last available measurement from a subject forward, replacing the entire sequence of missing values. A more sophisticated version predicts the next missing value using a regression relationship established from the available past data. These methods share the same drawbacks, although not all to the same degree. The data set that results will mimic a sample from the population of interest, itself determined by the analysis, only under particular and potentially unrealistic assumptions. Further, these assumptions depend critically on the missing value mechanism(s). For example, under certain missing value mechanisms, the process of imputation may recover the actual marginal behaviour required, whereas under other mechanisms, it may be wildly misleading. It is also under the simplest and most ignorable mechanisms that the relationship between imputation procedure and assumption is most easily deduced. In addition, the analysis of the completed data set will underestimate the true variability of the data.

One can therefore conclude that when there are missing values, simple methods of analysis do not necessarily imply simple or even accessible assumptions. Also, without properly understanding the assumptions being made in an analysis, adequate judge-

ment of validity of the analysis may not be possible.

4.9 Missing data mechanisms

In order to incorporate incompleteness or missingness into the modelling process, the nature of the missing value mechanism and its implications for statistical inference needs to be known. Rubin (1976) and Little and Rubin (1987, Chapter 6) make important distinctions between different missing value processes. These are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

4.9.1 Missing completely at random (MCAR)

Missingness is said to be MCAR if the missing values are independent of both unobserved and observed data. Therefore, cases with complete data are indistinguishable from cases with incomplete data. Heitjan (1997) provides an example of MCAR missing data: the case of a research associate shuffling raw data sheets and arbitrarily discarding some of the sheets.

4.9.2 Missing at random (MAR)

Missingness is said to be MAR if conditional on the observed data, the missingness is independent of the unobserved measurements. In other words, cases with incomplete data differ from cases with complete data, but the pattern of missing data is traceable or predictable from other variables in the database rather than from the specific variable on which the data is missing. An example of MAR is the following: suppose research participants with low-esteem are less likely to return for follow-up sessions in a study

that examines anxiety levels over time as a function of self-esteem, and the researcher measures self-esteem at the initial session. Then self-esteem can then be used to predict the missingness pattern of the incomplete data.

Another example of MAR is the following: suppose investigators administer a reading comprehension test at the beginning of a survey, then research participants with lower reading comprehension scores are less likely to complete the entire survey. In both examples, the actual variables for which data are missing are not the cause of the incomplete data. Instead, the cause of the missing data is due to some other external influence.

4.9.3 Missing not at random (MNAR)

In this case the pattern of data missingness is non-random and it is not predictable from other variables in the database. For example, if a participant in a weight-loss study does not attend a weigh-in due to concerns about his weight loss, his data are missing due to non-ignorable factors. In contrast to the MAR situation outlined above where data missingness is explainable by other measured variables in a study, non-ignorable missing data arise due to the data missingness pattern being explainable by the very variable(s) on which the data are missing.

Looking at these three missing data mechanisms further, assume that for subject i in the study, a sequence of measurements Y_{ij} is designed to be measured at occasions $j = 1, \dots, n_i$. Then the outcomes can be grouped into a vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})'$. In addition, for each occasion j , let's define the missing data indicators as:

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

These indicators can also be grouped into a vector \mathbf{R}_i which is of the same length as \mathbf{Y}_i .

The vector Y_i can then be partitioned into two sub-vectors such that Y_i^o is the vector containing those Y_{ij} for which $R_{ij} = 1$ and Y_i^m contains the remaining components. These sub-vectors are referred to as the observed and missing components respectively. Statistical modelling thus begins by considering the full data density

$$f(y_i, r_i | X_i, Z_i, \theta, \psi) = f(y_i | X_i, Z_i, \theta) f(r_i | y_i, X_i, \psi), \quad (4.15)$$

where X_i and Z_i are the design matrices for the fixed and random effects respectively; θ and ψ are vectors that parameterize the joint distribution; and r_i are the missing data indicators. Let $\theta = (\beta', \alpha')$ (fixed-effects and covariance parameters) and ψ describe the measurement and missingness processes respectively then the first factor of the factorization in 4.15 is the marginal density of the measurement process (measurement model) and the second one is the marginal density of the missingness process (missingness model). The difference in the missing data mechanisms is specified in the second factor of (4.15), that is

$$f(r_i | y_i, X_i, \psi) = f(r_i | y_i^o, y_i^m, X_i, \psi). \quad (4.16)$$

If (4.16) is independent of the measurements, that is if it assumes the form $f(r_i | X_i, \psi)$, then the process is termed MCAR. If (4.16) is independent of the unobserved (missing) measurements y_i^m , but depends on the observed measurements y_i^o , thereby assuming the form $f(r_i | y_i^o, X_i, \psi)$, then the process is termed MAR. Finally when (4.16) depends on the missing values y_i^m , the process is referred to as non-random missingness (MNAR). An MNAR is allowed to depend on y_i^o .

If interest is in the measurement model only, then the missingness model can be completely ignored. This means that under ignorability, MCAR and MAR have the same fitted measurement model. If missingness process is ignorable, then a valid analysis can be obtained through a likelihood-based analysis that ignores the missingness mechanism. This case, termed 'ignorable' by Rubin (1976), and Little and Rubin

(1987), leads to considerable simplification in the analysis (Diggle, 1989)

In most instances, the reasons for missingness are many and vary. It is therefore very difficult to justify the assumption of random missingness. Arguably, in the presence of non-random missingness, a wholly satisfactory analysis of the data is not feasible. One approach is to estimate from the available data the parameters of a model representing a non-random missingness mechanism. It may be difficult to justify the particular choice of missingness model, and it does not necessarily follow that data will contain information on the parameters of the particular model chosen. But where such information exists, the fitted model may provide some insight into the nature of the missingness process and into the sensitivity of the analysis to assumptions about this process. This is the route taken by Diggle and Kenward (1994); Diggle, Liang and Zeger (1994, Chapter 11); and Verbeke and Molenberghs (2000, Section 12.4). Further approaches are proposed by Laird, Lange and Stram (1987); Wu and Bailey (1989); Wu and Carroll (1988); and Greenless, Reece and Zieschang (1982). An overview of the different modelling approaches is given by Little (1995). Baker and Laird (1988); Stasny (1986); Conaway (1992); Park and Brown (1994); and Molenberghs, Kenward and Lesaffre (1997) look at cases of categorical outcomes. One feature, that is, however common in all the more complex approaches is that they rely on un-testable assumptions about the relation between the measurement process (often the primary interest) and the missingness process.

4.10 Fitting the longitudinal mixed model with missingness

The basic theory on which PROC MIXED is based holds even with unbalanced and missing data so long as the missing data is ignorable, since the PROC MIXED procedure allows routine fitting of ignorable models with the likelihood-based methods. From an interpretation given in Littell *et al.* (1996), the parameter estimates, standard errors and test statistics from a data set with ignorable missing data are not greatly different from those obtained using the complete data set. However, the degrees of freedom may be different. Therefore, before using the PROC MIXED procedure for analysis, one has to make sure that the ignorability assumption holds.

A general MAR test for multivariate data was provided by Simon and Simonoff (1986). Commenting on this test, Little (1988) proposes that when missing values are confined to a single variable y , the standard procedure is to compare the distribution of the fully observed variables for respondents and non-respondents, either informally or formally, via t tests for the differences in means.

In a regression setting, Simon and Simonoff (1986) explain how this test for MAR can be undertaken when only one independent variable has missing data while the other independent variables as well as the response variable have complete data. The model examined is the usual fixed-effects linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, with \mathbf{X} fixed and the error vector \mathbf{e} distributed as $N(\mathbf{0}, \sigma^2 \mathbf{I})$. Examined in detail is the case in which data is missing in one column of the \mathbf{X} matrix with the remaining columns complete. The vector \mathbf{y} is considered to be complete as well. Since the missingness in Durban-South data set is on the dependent variable \mathbf{y} , then this approach is not applicable to the problem at hand.

4.10.1 The logistic regression model MAR test

A logistic regression model was proposed by Diggle and Kenward (1994) to model the dropout process. A dropout in this case would mean that observations are made on a subject upto a certain point in time and thereafter no measurements are obtained. The logistic regression model can be used to explore the missing data process as follows (Fanta, 2003; and Verbeke and Molenberghs, 2000). Assume that the dropout probability at occasion j depends on both the current outcome Y_{ij} and the previous one Y_{ij-1} . This leads to the following model

$$\ln \left[\frac{P(R_{ij} = 0|y_i)}{1 - P(R_{ij} = 0|y_i)} \right] = \phi_0 + \phi_1 y_{ij} + \phi_2 y_{ij-1}, \quad (4.17)$$

$$\text{with } R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

The probability $P(R_{ij}=0|y_i)$ is the conditional probability of subject i dropping out at time j , depending on the current outcome Y_{ij} , and the previous one Y_{ij-1} . The assumption imposed here is that the relationship among the measurements on a subject are the same, even if some of these measurements are unobserved due to non-response. It is this assumption that allows one to infer about the informativeness of the process of missingness (Diggle and Kenward, 1994; Diggle *et al.*, 2002; and Verbeke and Molenberghs, 2000). The full model (4.17) represents the MNAR model where the process of missingness depends on the unobserved measurements. The reduced models with $\phi_1 = 0$ and $\phi_1 = \phi_2 = 0$ represent the MAR and MCAR models respectively. These three models are fitted and tested for significant difference. MAR missingness is assumed if the MNAR and MAR models are found not to be significantly different. MCAR missingness is assumed if MNAR and MCAR models are found not to be significantly different. Comparisons are made by using the LR test.

4.10.2 The MAR tests for random dropouts

Diggle (1989) and Ridout (1991) also provide testing procedures for random dropouts in repeated measurement data. Diggle (1989) provides a method for testing the hypothesis of random dropouts within groups, in the sense of being unrelated to measurement history, when the experiment has a completely randomized design.

In a reader reaction to Diggle (1989), Ridout (1991) pointed out the connection between Diggle's approach and a logistic regression analysis of the occurrence of dropouts. Ridout (1991) argued that, suppose there are r dropouts out of a total of R units that have not previously dropped out, then the test statistic proposed by Diggle would be the mean value of the covariate x for the r units that are about to drop out. This covariate is a function of the measurements taken before dropout occurs. However, Ridout (1991) conveniently considers the total of the covariate values, T say, rather than their mean. Under the null hypothesis, T is the total of a sample of size r drawn at random, without replacement, from the set of R values of the covariate x . It is this T statistic which is used in a conditional test of the hypothesis $\beta = 0$ in the logistic regression model

$$\text{logit}(\text{Pr}(\text{Dropout})) = \alpha + \beta x, \quad (4.18)$$

which provides a sensible reference distribution for calculating the significance level. Specifically the observed value of T is compared with other values that might have arisen for the given r dropouts. In addition, the conditional distribution of T does not depend on the nuisance parameter α . Ridout (1991) also alternatively considers the logistic regression model

$$\text{logit}(\text{Pr}(\text{Dropout})) = \alpha_i + \beta x, \quad (4.19)$$

where i indexes the different group-time combinations. The conditional test for $\beta = 0$ involves the statistic $T = \sum T_i$, evaluated conditionally on the number of dropouts

occurring in each group-time combination. The normal approximation for T is assumed and the test statistic along with the one-sided p -value (with 1 degree of freedom) is attained. Alternatively, the LR statistic for testing $\beta = 0$ referring to the chi-square distribution can be obtained which gives a two-tailed P -value. Equivalently, the signed square root of the LR statistic is asymptotically standard normal. Another alternative is to simply divide the maximum likelihood estimate of β by its standard error and this gives an asymptotically standard normal test statistic.

Finally, the hypothesis that the set of P -values (say p_u ($u = 1, \dots, t$)) obtained by applying the above procedures to each time point within each group is a random sample from a uniform distribution on $(0,1)$, is tested. As a general aid to interpretation, Diggle (1989) recommends the inspection of a probability plot of the p_u . If the dropouts are random, the p_u should behave like a random sample from a uniform distribution on $(0, 1)$. For a formal test, the one-sided or two-sided Kolmogorov-Smirnov statistic can be used. Note that for the final stage of the test, the independence of the P -values under the hypothesis of random dropouts within groups relies on the fact that once a unit drops out, it never returns. Thus, this test is also not applicable for intermittent missing values.

Chapter 5

Identification of the model

5.1 Preliminary analysis

As discussed in Chapter 2, there are five independent child-specific variables of which *School*, *Sex*, and *Asthma status* are categorical variables; and *Height* and *Weight* are continuous variables. In addition, the repeated SO_2 measurements, which are associated with the response variables, also independent variables. The response variable is the within-day variability ($WDVarFEV1$), which is also a repeated measurement. To simplify the interpretation of the results of the analysis, it is best to use the categorized *Height*, and *Weight* measurements.

5.1.1 Classification of the height and weight measurements

To classify the measurements of the *Height* and *Weight* variables, the National Center for Health Statistics/World Health Organization (NCHS/WHO) reference standards are used. According to Cogill (2003), the NCHS/WHO reference standards are most commonly used to standardize measurements. They were developed by the NCHS in the United States of America and are recommended for international use

by the WHO. The reference population chosen by NCHS was a statistically valid random population of healthy infants and children. The NCHS/WHO international reference tables have been used for standardizing anthropometric data around the world and can be found in the Food and Nutrition Technical Assistance (FANTA) website: www.fantaproject.org/publications/anthropom.shtml and the NCHS website: www.cdc.gov/nchs/about/major/nhanes/growthcharts. Anthropometry is the study of the technique of taking body measurements especially to assess and predict performance, health and survival of individuals and reflect the economic and social well being of the population.

These reference standards are used to standardize a child's measurement by comparing the child's measurement with the median or average measure for children of the same age and sex. For example, if the height of a nine year old boy is 130 cm, it is difficult to know if this is reflective of a healthy nine year old boy when there is no comparison to a reference standard. The reference or median height for a population of nine year old boys is 133.7 cm and the simple comparison of heights would conclude that the 130 cm tall child is almost 4 cm shorter than expected.

However the 4 cm height difference with the reference for a nine year old boy is not the same as a 4 cm height difference with the reference for a seventeen year old boy because of their relatively different body sizes due to their difference in age. Taking the age and sex into consideration, differences in measurements can be expressed in a number of ways namely (1) standard deviation units or Z-scores; (2) percentage of the median and (3) percentiles.

The Z-Score or standard deviation (SD) unit is defined as the difference between the value for an individual and the median value of the reference population with the same age or height as the individual, divided by the standard deviation of the values

of the reference population. This is,

$$\text{Z-score (or SD-score)} = \frac{\text{Observed value} - \text{median value of reference population}}{\text{Standard deviation of reference population}}$$

A positive Z-score means that the individual's measurement is higher than the reference median whereas a negative Z-score means that the individual's measurement is lower than the reference median. The distribution of Z-scores is a normal (bell-shaped or Gaussian) distribution (Cogill, 2003).

The percentage of the median is defined as the ratio of a measured or observed value of the individual to the median value of the reference population with the same age or height for the specific sex as the individual, expressed as a percentage. That is,

$$\text{Percentage of median} = \frac{\text{Observed value}}{\text{Median value of reference population}} \times 100.$$

If a child's measurement is exactly the same as the median of the reference population, then the child is said to be '100 of the median'. Lastly, the percentile is the rank position of an individual relative to the reference distribution, stated in terms of the percentage of the group the individual equals or exceeds.

Cut-offs are then used to enable the different individual measurements to be converted into prevalence statistics. Cut-offs are also used for identifying those children suffering from or at a higher risk of adverse outcomes. The most commonly used cut-off with Z-scores is -2 standard deviations, irrespective of the indicator used. This means that children with a Z-score below -2 SDs are considered to be moderately or severely malnourished for example children that are underweight (low weight-for-age), that are stunting (low height-for-age) or wasting (low weight-for-height).

The low weight-for-age index identifies the condition of being underweight for a specific age while a deficit in height-for-age is referred to as stunting. A low weight-for-height value helps to identify children suffering from current or acute undernutrition or wasting and is useful when exact ages are difficult to determine.

The cut-off points for different malnutrition classification systems are listed in Table 5.1. The most widely used system is the WHO classification (Z-scores). The Road-to-Health (RTH) system is typically seen in clinic-based growth-monitoring systems. The Gomez system was widely used in the 1960s and 1970s, but is only used in a few countries now. Mild, moderate and severe are different in each of the classification systems listed in Table 5.1. It is important that the same system is used to analyse and to present the data. The RTH and Gomez classification systems typically use weight-for-age.

Table 5.1: Malnutrition Classification Systems

System	Cut-off	Malnutrition classification
WHO	< -1 to > -2 Z-score	mild
	< -2 to > -3 Z-score	moderate
	< -3 Z-score	severe
RTH	> 80% of median	normal
	60% - < 80% of median	mild-to-moderate
	< 60% of median	severe
GOMEZ	> 90% of median	normal
	75% - < 90% of median	mild
	60% - < 75% of median	moderate
	< 60% of median	severe

Source: <http://www.fantaproject.org/publications/anthropom.shtml>

For this study, the Z-scores are used to classify the *Weight* and *Height* measurements. The *Weight* variable is coded as follows: 1 for Risk of underweight (< -1.25 Z-score); 2 for Normal (-1.25 ≤ Z-score ≤ 1.25); and 3 for Risk of overweight (> 1.25 Z-score). The *Height* variable is coded as follows: 1 for Short (< -1.25 Z-score); 2 for Normal (-1.25 ≤ Z-score ≤ 1.25); and 3 for Tall (> 1.25 Z-score)

5.1.2 The distribution of the children by the child-specific variables

In this section, the structural distribution of the five child-specific variables *School*, *Sex*, *Asthma status* and the classified variables *Height* and *Weight* is explored. A total of 233 children are considered for the analysis. Of these, 58 are from Nizam Road, 46 are from Assegai, 16 are from Dirkie Uys, 59 are from Ferndale and 54 are from Ngazana as shown in Table 5.2. Dirkie Uys is represented by the least number of children which was due to high refusal rate of the parents and guardians for their children to participate in the study. Table 5.2 also shows that 58% of the 233 children are female while 24% of the total are persistent asthmatics and 27% are asthmatics.

Table 5.2: The distribution of children by sex, asthma status and school

School	Sex			Asthma status			
	Male	Female	Total	Persistent Asthmatics	Asthmatics	Normal	Total
Nizam Road	29	29	58	15	15	28	58
Assegai	17	29	46	17	9	20	46
Dirkie Uys	9	7	16	3	7	6	16
Ferndale	27	32	59	10	17	32	59
Ngazana	17	37	54	10	15	29	54
Total	99	134	233	55	63	115	233

Table 5.3 shows the distribution of children by *Height* and *Weight*. About 16% of the children in the study are too short for their age while 5.6% are too tall for their age. In addition, 18% are underweight while 15% are overweight.

Table 5.3: The frequency distribution of children by height, weight and school

School	Height-for-age classification				Weight-for-age classification			
	Short-for-age	Normal-for-age	Tall-for-age	Total	Under-weight	Normal-for-age	Over-weight	Total
Nizam Road	17	40	1	58	27	24	7	58
Assegai	5	35	6	46	4	35	7	46
Dirkie Uys	1	14	1	16		12	4	16
Ferndale	4	53	2	59	5	45	9	59
Ngazana	10	41	3	54	6	40	8	54
Total	37	183	13	233	42	156	35	233

5.1.3 The data structure of the SO₂ pollutant measures

In this section, the pollutant measures are explored. Figure 5.1 and Figure 5.2 present the pollutant measures for each school and for the Phase I blow exercise which was carried out between the 31st May 2004 and 18th June 2004. The summary measurement displayed in Figure 5.1 are the peak counts for the 24 hours prior to the child's first blow of the day, while the 8-hour maximum of the moving averages computed 24 hours prior to the child's first blow of the day are represented in Figure 5.2.

As shown in both graphs, Nizam road, Assegai and Dirkie Uys schools recorded more SO₂ activity when compared to Ferndale and Ngazana schools. Figure 5.1 shows that Nizam road registered the highest total number of times that the pollutant levels went beyond the standard guidelines, with 24 peaks. This is followed by Assegai, with 19 peaks. Ngazana had the lowest total number of peaks, with 6 peaks. Dirkie Uys and Ferndale schools had 8 and 11 total number of peaks respectively. The actual peak counts are shown in Appendix A.1 while the SO₂ readings that are above the standard guideline of 191ppb are shown in Appendix A.3. Despite Ngazana having the lowest number of peaks, it registered the highest pollutant levels followed by Ferndale,

as shown in Appendix A.3.

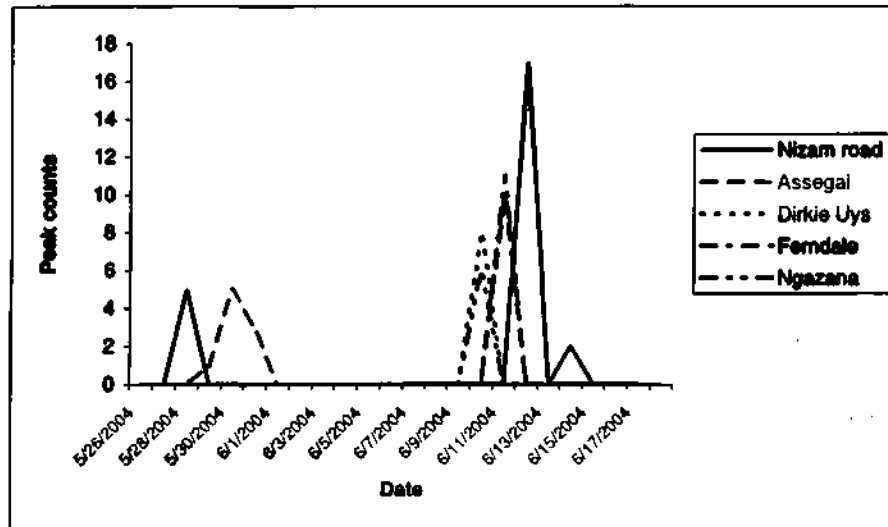


Figure 5.1: The graphical representation of the peak counts.

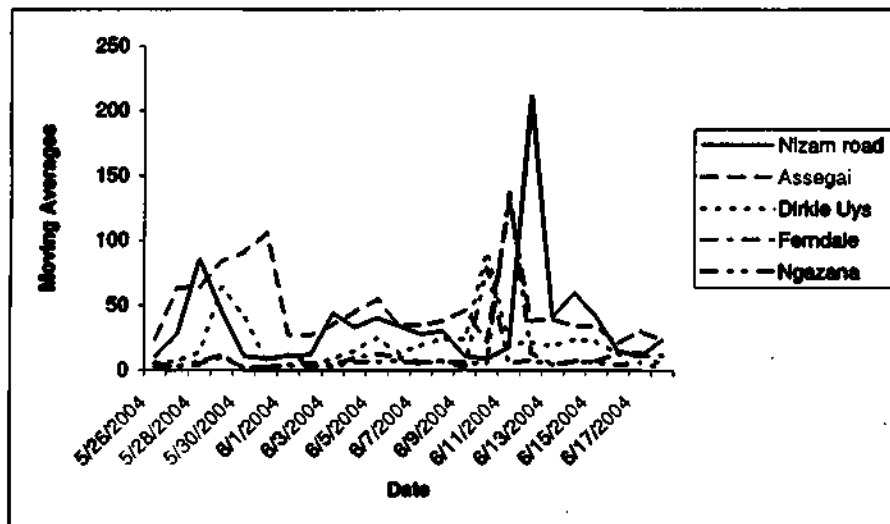


Figure 5.2: The graphical representation of the 8-hour maximum of the SO_2 moving averages.

For the alternative pollutant measure, which is the 8-hour maximum of the moving averages, a graphical representation of all the 8-hour maximum values for each school on a particular date are shown in Figure 5.2. These are generated from the table in Appendix A.2. Figure 5.2 shows Nizam road and Assegai schools with the highest

values while Ngazana registered the lowest values with one hike on the 10th of June 2004. The same conclusion can be drawn from the average and total of the 8-hour maximum values generated per school (Appendix A.2).

To account for the possible lag effects 1 - 5 day-lag pollutant variables are considered for each of the two summary measurements as described in Section 2.2. This results in a total of ten pollutant variables (five peaks and five 8-hour maximum variables) to be included in the model. However, before doing so, they are first tested for multicollinearity. This is first done by using their correlation coefficients shown in Table 5.4.

Variables *Peak1* to *Peak5* are the 1 - 5 day-lag peaks pollutant variables while *HrMax1* to *HrMax5* represent the 1 - 5 day-lag 8-hour maximum pollutant variables. Most of the pollutant variables are significantly correlated at 5% significance level. Further more, the correlation between the peak and 8-hour maximum variables for the same lag period are large. That is to say, *Peak1* is highly correlated with the *HrMax1* with correlation coefficient 0.85816, *Peak2* is highly correlated with the *HrMax2* with correlation coefficient 0.74655, *Peak3* is highly correlated with the *HrMax3* with correlation coefficient 0.85819, *Peak4* is highly correlated with the *HrMax4* with correlation coefficient 0.88096, and *Peak5* is highly correlated with the *HrMax5* with correlation coefficient 0.83478.

Table 5.4: Pearson Correlation coefficients and prob > |r| under $H_0: \rho = 0$

	Peak1	Peak2	Peak3	Peak4	Peak5	HrMax1	HrMax2	HrMax3	HrMax4	HrMax5
Peak1	1	0.02107	0.08337	-0.07164	-0.08304	0.85816	-0.00398	0.10327	-0.06256	-0.09175
		0.2289	<.0001	<.0001	<.0001	<.0001	0.8202	<.0001	0.0003	<.0001
Peak2	0.02107	1	0.05799	0.13168	-0.08571	0.14936	0.74855	0.14214	0.20534	0.01059
	0.2289		0.0009	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.5453
Peak3	0.08337	0.05799	1	-0.0063	-0.05718	0.19669	0.27646	0.85819	0.03917	-0.01714
	<.0001	0.0009		0.7192	0.0011	<.0001	<.0001	<.0001	0.0253	0.3277
Peak4	-0.07164	0.13168	-0.0063	1	-0.05394	0.02217	0.17446	0.13816	0.88096	-0.01892
	<.0001	<.0001	0.7192		0.0021	0.2056	<.0001	<.0001	<.0001	0.2799
Peak5	-0.09304	-0.08571	-0.05718	-0.05394	1	-0.00778	0.00229	-0.01349	0.05307	0.83478
	<.0001	<.0001	0.0011	0.0021		0.8571	0.896	0.4411	0.0024	<.0001
HrMax1	0.85816	0.14936	0.19669	0.02217	-0.00778	1	0.33236	0.36109	0.17087	0.14697
	<.0001	<.0001	<.0001	0.2056	0.8571		<.0001	<.0001	<.0001	<.0001
HrMax2	-0.00398	0.74855	0.27646	0.17446	0.00229	0.33236	1	0.51485	0.41953	0.27879
	0.8202	<.0001	<.0001	<.0001	0.896	<.0001		<.0001	<.0001	<.0001
HrMax3	0.10327	0.14214	0.85819	0.13816	-0.01349	0.36109	0.51485	1	0.33003	0.21201
	<.0001	<.0001	<.0001	<.0001	0.4411	<.0001	<.0001		<.0001	<.0001
HrMax4	-0.06256	0.20534	0.03917	0.88096	0.05307	0.17087	0.41953	0.33003	1	0.26563
	0.0003	<.0001	0.0253	<.0001	0.0024	<.0001	<.0001	<.0001		<.0001
HrMax5	-0.09175	0.01059	-0.01714	-0.01892	0.83478	0.14697	0.27879	0.21201	0.26563	1
	<.0001	0.5453	0.3277	0.2799	<.0001	<.0001	<.0001	<.0001	<.0001	

The alternative indicator of multicollinearity known as the variance inflation factor (VIF) is used to detect multicollinearity to the pollutant measures. As shown in Table 5.5, all the VIF values are larger than one. However, a serious problem of multicollinearity is detected for variables *Peak4*, *HrMax3* and *HrMax4* as their VIF values are greater than 10.

Table 5.5: The Variance inflation factor for the 10 pollutant independent variables

Variable (j)	Variance Inflation Factor (VIF)
Peak1	8.51640
Peak2	4.25317
Peak3	8.27524
Peak4	11.31774
Peak5	6.20494
HrMax1	10.07748
HrMax2	7.92122
HrMax3	12.02757
HrMax4	15.24270
HrMax5	9.61287

To solve the multicollinearity problem, the PCA is applied to the ten pollutant variables using the SAS PROC FACTOR procedure. The PCA is based on the correlation matrix of the pollutant variables given in Table 5.4. The eigenvalues of the pollutant correlation matrix are found to be 2.96853292, 1.98310222, ..., 0.03028539 as shown in Table 5.6. These add up to 10, the sum of the diagonal terms in the correlation matrix. The corresponding eigenvectors are shown in Table 5.7 and are standardized so that the sum of the squares of the components is unity for each one of them. These eigenvectors provide the coefficients of the standardized pollutant variables in the principal components.

The eigenvalue for a principal component indicates the variance that it accounts for out of the total variances of 10. Thus, the first principal component (Z_1) accounts for $(2.96853292/10)100\% = (0.2969 \times 100)\% = 29.69\%$ of the total variance, the second for 19.83%, the third for 17.76% and so on as shown in the column "Proportion" in Table 5.6. Clearly, the first component is more important than the others, the second is more important than the third and so on. Notice also that the first five principal

components (eigenvalues > 1) account for 94.5% of the total variance. The eigenvectors after rotation are shown in Table 5.8.

Table 5.6: The eigenvalues of the correlation matrix for the 10 pollutant variables

Component (Z)	Eigenvalue	Proportion	Cumulative
1	2.968533	0.2969	0.297
2	1.983102	0.1983	0.495
3	1.776366	0.1776	0.673
4	1.454928	0.1455	0.818
5	1.264508	0.1265	0.945
6	0.334569	0.0335	0.978
7	0.094515	0.0095	0.988
8	0.050307	0.0050	0.993
9	0.042888	0.0043	0.997
10	0.030285	0.0030	1.000

Table 5.7: The eigenvectors of the correlation matrix for the 10 pollutant variables

Z	Eigenvector, coefficient of									
	Peak1	Peak2	Peak3	Peak4	Peak5	HrMax1	HrMax2	HrMax3	HrMax4	HrMax5
1	0.25955	0.51829	0.54183	0.44251	0.09756	0.57286	0.80182	0.77988	0.65610	0.35194
2	-0.68350	0.01850	-0.33755	0.39553	0.59726	-0.53271	0.07728	-0.19066	0.45753	0.59419
3	0.30158	-0.26121	0.09191	-0.56738	0.73295	0.33553	-0.09420	0.10078	-0.36151	0.68736
4	0.57364	-0.00956	-0.64077	0.35874	0.06103	0.48841	-0.12464	-0.47336	0.31917	0.05245
5	0.07036	-0.75762	0.33327	0.38609	0.02046	0.02292	-0.50237	0.29400	0.29140	-0.02787
6	0.13921	0.26681	0.21941	0.18644	0.27253	-0.10265	-0.19257	-0.09494	-0.09681	-0.14294
7	0.04480	0.12484	-0.03218	-0.05696	-0.10798	-0.05636	-0.17613	0.08287	0.06289	0.12635
8	0.13733	-0.04372	0.00616	-0.01434	-0.01071	-0.14850	0.07773	-0.01191	0.02316	0.01937
9	-0.01456	-0.00459	0.11174	0.00632	-0.06910	0.02418	0.00523	-0.13348	0.00997	0.08282
10	0.00684	-0.00487	-0.02556	0.09911	-0.03983	-0.00874	0.01273	0.03056	-0.11626	0.05906

Table 5.8 shows that Z_1 (for component 1) is high if *Peak3* and *HrMax3* are high. The other coefficients are very low which means that their values do not affect Z_1 . The same applies to the other principal components and as shown by the bold figures in Table 5.8. Since the first five principal components account for 94.5% of the total

variance, the other five principal components (that account for a small proportion of the variation in the data) can be discarded.

Table 5.8: The rotated eigenvectors of the correlation matrix for the 10 pollutant variables

Z	Rotated eigenvector, coefficient of									
	Peak1	HrMax1	Peak2	HrMax2	Peak3	HrMax3	Peak4	HrMax4	Peak5	HrMax5
1	0.01638	0.17155	0.00837	0.29187	0.97934	0.92276	0.00935	0.10055	-0.03457	0.06030
2	-0.06071	0.06602	0.07551	0.18564	-0.03670	0.16232	0.98073	0.94204	-0.01897	0.07533
3	-0.08116	0.06964	-0.04803	0.11607	-0.04674	0.07860	-0.06337	0.12888	0.96729	0.94596
4	0.97490	0.93997	0.03606	0.08761	0.05060	0.13838	-0.02963	0.03127	-0.03658	0.01878
5	-0.02601	0.13133	0.98861	0.79620	0.04126	0.14145	0.04744	0.15705	-0.04915	0.06870
6	-0.09990	0.14384	-0.11486	0.47121	-0.05783	0.15128	-0.06690	0.13481	-0.09120	0.14354
7	-0.03017	0.04315	-0.00770	0.04250	-0.05917	0.08868	-0.07490	0.10217	-0.22015	0.25544
8	-0.01920	0.03024	-0.00330	0.02498	-0.15996	0.20993	-0.03512	0.05607	-0.03765	0.05251
9	-0.16431	0.19200	-0.00240	0.02296	-0.01341	0.02851	-0.01066	0.02251	-0.01548	0.02735
10	-0.00972	0.01565	-0.00190	0.01662	-0.02190	0.03473	-0.13965	0.16716	-0.02474	0.03767

Notice that the principal components combine the variables for the same lag period, for example, *Peak3* and *HrMax3* are combined in the Z_1 , *Peak4* and *HrMax4* are combined in the Z_2 and so on. This limits the separate determination of the effect of the two summary measures (that is the peaks and the 8-hour maximum measures). As mentioned in Section 2.2, the peaks and the 8-hour maximum pollutant measures represent two different measures of the pollutant, therefore they are both useful in the analysis.

To enable the separate determination of the effect of each of the pollutant measures, two summary measures (peaks and 8-hour maximum) are split such that an analysis of the five variables for each summary measure is done separately. That is to say, have two models, the first one with the peak variables *Peak1* to *Peak5* as the pollutant independent variables and the second one with the 8-hour maximum variables *HrMax1* to *HrMax5* as the pollutant independent variables.

Table 5.9: The Variance inflation factor for the pollutant measures with the peaks and the 8-hour maximum computed separately

Peaks		8-hour maximum	
Variable (j)	Variance Inflation Factor (VIF _j)	Variable (j)	Variance Inflation Factor (VIF _j)
Peak1	1.02214	HrMax1	1.19418
Peak2	1.02517	HrMax2	1.64405
Peak3	1.01235	HrMax3	1.51251
Peak4	1.02278	HrMax4	1.28202
Peak5	1.02208	HrMax5	1.14945

Table 5.4 shows that the correlations among the peaks variables, and among the 8-hour maximum variables are low in absolute value but are significantly different from zero. However, the VIF values presented in Table 5.9 show that when the peaks and 8-hour maximum measures are fitted separately, all the respective VIF values are close to one. This shows that there are weak linear relationships between variables *Peak1*, ..., *Peak5* as well as among the variables *HrMax1*, ..., *HrMax5*. The significance of the correlations among the five peak variables and among the five 8-hour maximum pollutant variables can be attributed to the large size of the data. In conclusion, it was decided to fit two separate models, one with the five peaks pollutant measures as the pollutant variables and the other with the five 8-hour maximum pollutant measures as the pollutant variables.

5.1.4 The fixed-effects structure

In order to choose a fixed-effects structure for the longitudinal mixed model to be fitted, a profile of how relevant sub-populations evolve over time needs to be described. This aids one to visualize the pattern of the data over the different times (time being expressed by days in the study). The individual profiles for 30 randomly selected

children for each of the three asthma status categories are displayed in Figure 5.3. The mean profiles for each asthma status of all the children in the study are shown in Figure 5.4.

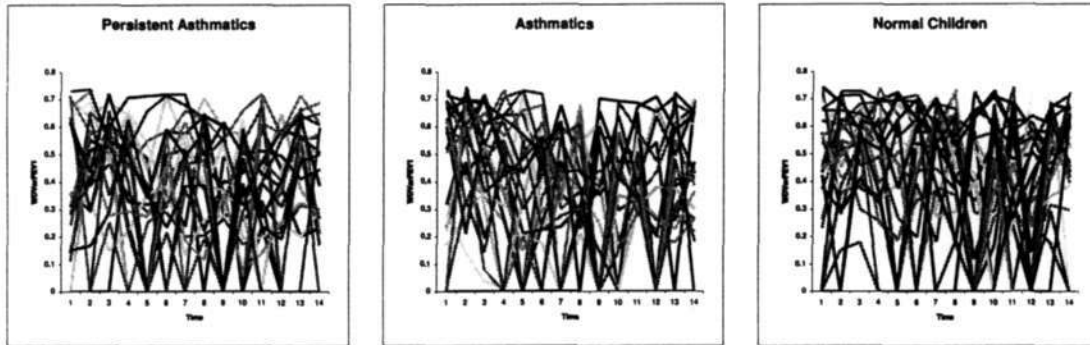


Figure 5.3: Individual profiles of 30 randomly selected children in the study.

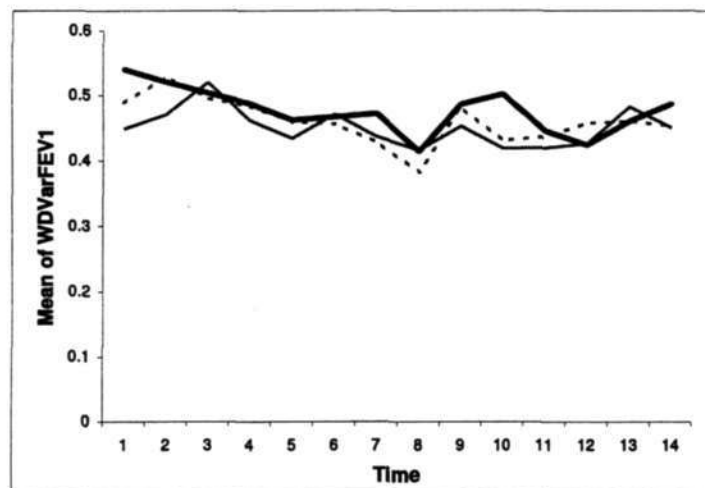


Figure 5.4: Mean profiles for the WVarFEV1 response variable. The solid line represents the persistent asthmatics, the dashed line the asthmatics and the bold line the normal children.

By looking at the profile plots in Figure 5.3 and Figure 5.4, it is difficult to determine the most appropriate functional form of dependence of the summary FEV1 measure on time. The scatter-plot smoother described in Section 4.1 is therefore used to find the functional dependence without imposing parametric assumptions about the dependence. Using the SAS PROC LOESS procedure, the non-parametric curves dis-

played in Figure 5.5 are constructed. This loess fit is applied to the randomly selected 30 children for each asthma status sub-population.

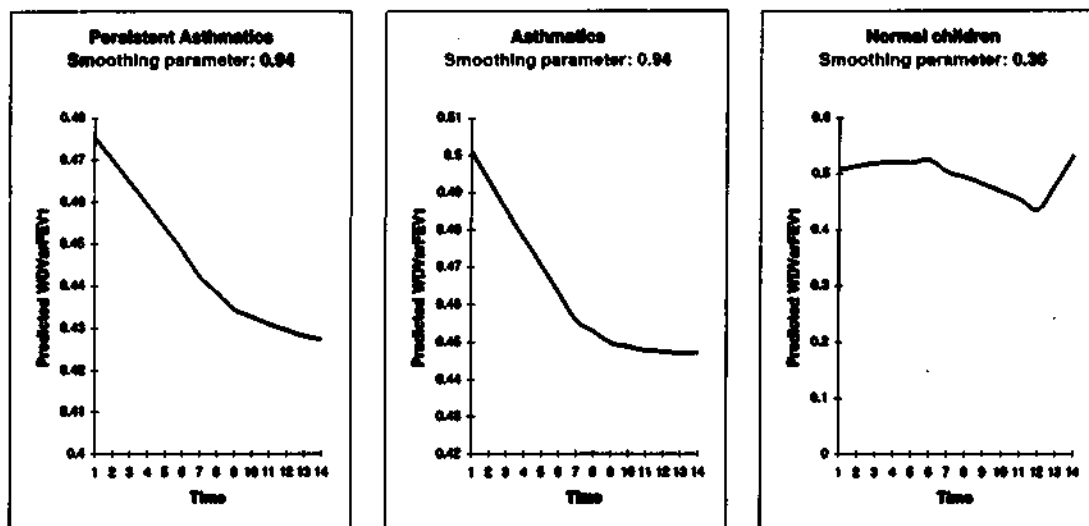


Figure 5.5: The Loess smooth curves of WdVarFEV1 against day for the random sample of respective asthma status sub-populations. The smoothing parameter is chosen using the AIC criteria.

Figure 5.5 shows that persistent asthmatics experience a sharp fall in the WdVarFEV1 at the beginning of the phase, thereafter the decrease slows down. These curves suggest a curvilinear statistical relationship between the WdVarFEV1 and time. Asthmatics experience a sharp fall in WdVarFEV1 at the beginning, reaching a minimum in the middle of the phase where it stabilizes up to the end of the phase. Asthmatic curves, therefore, also suggest a curvilinear statistical relationship between the WdVarFEV1 and day. As for the normal children, there is approximately no change in WdVarFEV1 throughout the phase except for a slight fall, then rise towards the end of the phase. In general, the curve for the normal children shows approximately no relationship between the WdVarFEV1 and time. The patterns in Figure 5.5 therefore suggest that it would be reasonable to approximate the relationship by a quadratic function of time.

5.2 Selection of a preliminary mean and random-effects structure

The plot of individual profiles in Figure 5.3, mean profiles in Figure 5.4 and the loess smoothed curves in Figure 5.5 suggest that a polynomial of degree two seems adequate to explain the $WDVarFEV1$ as a function of time. The preliminary informal analysis also indicates that the intercept and slopes are different for the different children. This leads to the inclusion of the child-specific regression coefficients, that represent the random effect, in the longitudinal mixed model. The first stage model can therefore be presented as a quadratic function over time, where the day is expressed as time of the model as follows:

$$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \beta_{2i}t_{ij}^2 + \beta_1Peak1_{ij} \\ + \beta_2Peak2_{ij} + \beta_3Peak3_{ij} + \beta_4Peak4_{ij} + \beta_5Peak5_{ij} + \epsilon_{ij},$$

where $i=1, \dots, 233$ refers to the i -th child and $j=1, \dots, n_i$ refers to the j -th repeated measurements for the i -th child. The $Peak1, \dots, Peak5$ are the SO_2 peak count variables for the five day lags. In the alternative model, the SO_2 variables considered are the 8-hour maximum measures denoted by $HrMax1, \dots, HrMax5$. The time and squared time points are represented by t_{ij} and t_{ij}^2 respectively.

In the second stage of the model development, the child-specific intercept (β_{0i}) and slopes (β_{1i}, β_{2i}) are related to the child-specific categorical variables *School*, *Sex*, *Asthma status*, *Height*, and *Weight*. The model at the second stage then becomes:

$$\begin{aligned}
\beta_{0i} = & \gamma_{01}\text{School}_{1i} + \gamma_{02}\text{School}_{2i} + \gamma_{03}\text{School}_{3i} \\
& + \gamma_{04}\text{School}_{4i} + \gamma_{05}\text{School}_{5i} \\
& + \delta_{01}\text{Sex}_{1i} + \delta_{02}\text{Sex}_{2i} \\
& + \theta_{01}\text{AsthmaStatus}_{1i} + \theta_{02}\text{AsthmaStatus}_{2i} + \theta_{03}\text{AsthmaStatus}_{3i} \\
& + \lambda_{01}\text{Weight}_{1i} + \lambda_{02}\text{Weight}_{2i} + \lambda_{03}\text{Weight}_{3i} \\
& + \alpha_{01}\text{Height}_{1i} + \alpha_{02}\text{Height}_{2i} + \alpha_{03}\text{Height}_{3i} + b_{0i},
\end{aligned}$$

$$\begin{aligned}
\beta_{1i} = & \gamma_{11}\text{School}_{1i} + \gamma_{12}\text{School}_{2i} + \gamma_{13}\text{School}_{3i} \\
& + \gamma_{14}\text{School}_{4i} + \gamma_{15}\text{School}_{5i} \\
& + \delta_{11}\text{Sex}_{1i} + \delta_{12}\text{Sex}_{2i} \\
& + \theta_{11}\text{AsthmaStatus}_{1i} + \theta_{12}\text{AsthmaStatus}_{2i} + \theta_{13}\text{AsthmaStatus}_{3i} \\
& + \lambda_{11}\text{Weight}_{1i} + \lambda_{12}\text{Weight}_{2i} + \lambda_{13}\text{Weight}_{3i} \\
& + \alpha_{11}\text{Height}_{1i} + \alpha_{12}\text{Height}_{2i} + \alpha_{13}\text{Height}_{3i} + b_{1i},
\end{aligned}$$

$$\begin{aligned}
\beta_{2i} = & \gamma_{21}\text{School}_{1i} + \gamma_{22}\text{School}_{2i} + \gamma_{23}\text{School}_{3i} \\
& + \gamma_{24}\text{School}_{4i} + \gamma_{25}\text{School}_{5i} \\
& + \delta_{21}\text{Sex}_{1i} + \delta_{22}\text{Sex}_{2i} \\
& + \theta_{21}\text{AsthmaStatus}_{1i} + \theta_{22}\text{AsthmaStatus}_{2i} + \theta_{23}\text{AsthmaStatus}_{3i} \\
& + \lambda_{21}\text{Weight}_{1i} + \lambda_{22}\text{Weight}_{2i} + \lambda_{23}\text{Weight}_{3i} \\
& + \alpha_{21}\text{Height}_{1i} + \alpha_{22}\text{Height}_{2i} + \alpha_{23}\text{Height}_{3i} + b_{2i},
\end{aligned}$$

where *Nizam Road*, *Assegai*, *Dirkie Uys*, *Ferndale* and *Ngazana* are the school factor levels; *Male* and *Female* are the sex factor levels; *Persistent asthmatic*, *Asthmatic* and *Normal* are the asthma status factor levels; *Underweight*, *Normal weight* and

Overweight are the weight factor levels; and *Short*, *Normal height* and *Tall* are the height factor levels. These second stage models can be rewritten as follows:

$$\beta_{0i} = \sum_{c=1}^5 \gamma_{0c} \text{School}_{ci} + \sum_{s=1}^2 \delta_{0s} \text{Sex}_{si} \\ + \sum_{u=1}^3 \theta_{0u} \text{AsthmaStatus}_{ui} + \sum_{w=1}^3 \lambda_{0w} \text{Weight}_{wi} + \sum_{h=1}^3 \alpha_{0h} \text{Height}_{hi} + b_{0i},$$

where $c=1, \dots, 5$ corresponds to *Nizam Road*, *Assegai*, *Dirkie Uys*, *Ferndale* and *Ngazana* respectively; $s=1, 2$ corresponds to *Male* and *Female* respectively; $u=1, 2, 3$ corresponds to *Persistent asthmatic*, *Asthmatic* and *Normal* respectively; $w=1, 2, 3$ corresponds to *Underweight*, *Normal weight* and *Overweight* respectively; and $h=1, 2, 3$ corresponds to *Short*, *Normal height* and *Tall* respectively. Likewise,

$$\beta_{1i} = \sum_{c=1}^5 \gamma_{1c} \text{School}_{ci} + \sum_{s=1}^2 \delta_{1s} \text{Sex}_{si} \\ + \sum_{u=1}^3 \theta_{1u} \text{AsthmaStatus}_{ui} + \sum_{w=1}^3 \lambda_{1w} \text{Weight}_{wi} + \sum_{h=1}^3 \alpha_{1h} \text{Height}_{hi} + b_{1i},$$

and

$$\beta_{2i} = \sum_{c=1}^5 \gamma_{2c} \text{School}_{ci} + \sum_{s=1}^2 \delta_{2s} \text{Sex}_{si} \\ + \sum_{u=1}^3 \theta_{2u} \text{AsthmaStatus}_{ui} + \sum_{w=1}^3 \lambda_{2w} \text{Weight}_{wi} + \sum_{h=1}^3 \alpha_{2h} \text{Height}_{hi} + b_{2i}.$$

The combined model in this case is

$$\begin{aligned}
y_{ij} = & \sum_{C=1}^5 \gamma_{0c} \text{School}_{ci} + \sum_{S=1}^2 \delta_{0s} \text{Sex}_{si} \\
& + \sum_{U=1}^3 \theta_{0u} \text{AsthmaStatus}_{ui} + \sum_{W=1}^3 \lambda_{0w} \text{Weight}_{wi} + \sum_{H=1}^3 \alpha_{0h} \text{Height}_{hi} \\
& + \sum_{C=1}^5 \gamma_{1c} \text{School}_{ci} t_{ij} + \sum_{S=1}^2 \delta_{1s} \text{Sex}_{si} t_{ij} \\
& + \sum_{U=1}^3 \theta_{1u} \text{AsthmaStatus}_{ui} t_{ij} + \sum_{W=1}^3 \lambda_{1w} \text{Weight}_{wi} t_{ij} + \sum_{H=1}^3 \alpha_{1h} \text{Height}_{hi} t_{ij} \\
& + \sum_{C=1}^5 \gamma_{2c} \text{School}_{ci} t_{ij}^2 + \sum_{S=1}^2 \delta_{2s} \text{Sex}_{si} t_{ij}^2 \\
& + \sum_{U=1}^3 \theta_{2u} \text{AsthmaStatus}_{ui} t_{ij}^2 + \sum_{W=1}^3 \lambda_{2w} \text{Weight}_{wi} t_{ij}^2 + \sum_{H=1}^3 \alpha_{2h} \text{Height}_{hi} t_{ij}^2 \\
& + \beta_1 \text{Peak1}_{ij} + \beta_2 \text{Peak2}_{ij} + \beta_3 \text{Peak3}_{ij} + \beta_4 \text{Peak4}_{ij} + \beta_5 \text{Peak5}_{ij} \\
& + b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \varepsilon_{ij}.
\end{aligned}$$

Let \mathbf{y}_i be the n_i -dimensional vector of all the repeated measurements for the i -th child, that is $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})'$, then the model can be summarized as follows:

$$\begin{aligned}
\mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i \\
\mathbf{b}_i &\sim N(\mathbf{0}, \mathbf{D}) \\
\boldsymbol{\varepsilon}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_i) \\
\mathbf{y}_i &\sim N(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \boldsymbol{\Sigma}_i),
\end{aligned}$$

where the fixed effects parameter $\boldsymbol{\beta} = (\gamma_{01}, \dots, \gamma_{05}, \delta_{01}, \delta_{02}, \theta_{01}, \theta_{02}, \theta_{03}, \lambda_{01}, \lambda_{02}, \lambda_{03}, \alpha_{01}, \alpha_{02}, \alpha_{03}, \gamma_{11}, \dots, \gamma_{15}, \delta_{11}, \delta_{12}, \theta_{11}, \theta_{12}, \theta_{13}, \lambda_{11}, \lambda_{12}, \lambda_{13}, \alpha_{11}, \alpha_{12}, \alpha_{13}, \gamma_{21}, \dots, \gamma_{25}, \delta_{21}, \delta_{22}, \theta_{21}, \theta_{22}, \theta_{23}, \lambda_{21}, \lambda_{22}, \lambda_{23}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \beta_1, \dots, \beta_5)'$ and the child-specific effects \mathbf{b}_i are represented by $\mathbf{b}_i = (b_{0i}, b_{1i}, b_{2i})'$.

Random effects $\mathbf{b}_1, \dots, \mathbf{b}_N, \boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_N$ are assumed to be independent with \mathbf{D} as a general (3×3) covariance matrix with (i, j) element $d_{ij} = d_{ji}$ and $\boldsymbol{\Sigma}_i$ as an

$(n_i \times n_i)$ covariance matrix respectively. A somewhat less crucial assumption is that these random effects are drawn from a normally distributed population. These child-specific effects will usually not be independent but will be correlated. However, it is assumed that for different children, the pairs of the random effects, (b_{0i}, b_{1i}, b_{2i}) , are independent. The y_i is thus also assumed to be normally distributed with mean vector $X_i\beta$ and covariance matrix $V_i = Z_i D Z_i' + \Sigma_i$.

The model with the peaks as the independent pollutant measures will be referred to as the 'peaks pollutant model' while the model with the 8-hour maximum as independent pollutant measures will be referred to as the '8-hour maximum pollutant model'. The peaks pollutant model will be discussed in Chapter 6 and the 8-hour maximum pollutant model will be dealt with in Chapter 7.

Chapter 6

Analysis of the peaks pollutant model

The peaks pollutant model for the j -th repeated measurement of the i -th child can be represented by

$$\begin{aligned}
 y_{ij} = & \sum_{c=1}^5 \gamma_{0c} \text{School}_{ci} + \sum_{s=1}^2 \delta_{0s} \text{Sex}_{si} + \sum_{u=1}^3 \theta_{0u} \text{AsthmaStatus}_{ui} + \sum_{w=1}^3 \lambda_{0w} \text{Weight}_{wi} \\
 & + \sum_{h=1}^3 \alpha_{0h} \text{Height}_{hi} + \sum_{c=1}^5 \gamma_{1c} \text{School}_{ci} t_{ij} + \sum_{s=1}^2 \delta_{1s} \text{Sex}_{si} t_{ij} + \sum_{u=1}^3 \theta_{1u} \text{AsthmaStatus}_{ui} t_{ij} \\
 & + \sum_{w=1}^3 \lambda_{1w} \text{Weight}_{wi} t_{ij} + \sum_{h=1}^3 \alpha_{1h} \text{Height}_{hi} t_{ij} + \sum_{c=1}^5 \gamma_{2c} \text{School}_{ci} t_{ij}^2 + \sum_{s=1}^2 \delta_{2s} \text{Sex}_{si} t_{ij}^2 \\
 & + \sum_{u=1}^3 \theta_{1u} \text{AsthmaStatus}_{ui} t_{ij}^2 + \sum_{w=1}^3 \lambda_{2w} \text{Weight}_{wi} t_{ij}^2 + \sum_{h=1}^3 \alpha_{2h} \text{Height}_{hi} t_{ij}^2 \\
 & + \beta_1 \text{Peak1}_{ij} + \beta_2 \text{Peak2}_{ij} + \beta_3 \text{Peak3}_{ij} + \beta_4 \text{Peak4}_{ij} + \beta_5 \text{Peak5}_{ij} \\
 & + b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \epsilon_{ij},
 \end{aligned} \tag{6.1}$$

where $i=1, \dots, 233$ are children in the study; $j=1, \dots, n_i$ are the repeated measurements; $c=1, \dots, 5$ corresponds to *Nizam Road, Assegai, Dirkie Uys, Ferndale* and

Ngazana respectively; $s=1, 2$ corresponds to *Male* and *Female* respectively; $u=1, 2, 3$ corresponds to *Persistent asthmatic*, *Asthmatic* and *Normal* respectively; $w=1, 2, 3$ corresponds to *Underweight*, *Normal weight* and *Overweight* respectively; and $h=1, 2, 3$ corresponds to *Short*, *Normal height* and *Tall* respectively. The dependent variable *WdVarFEV1* is represented by y_{ij} ; while the pollutant variables are the repeated measurements $Peak1_{ij}$, $Peak2_{ij}$, $Peak3_{ij}$, $Peak4_{ij}$ and $Peak5_{ij}$. The time and squared time points are represented by t_{ij} and t_{ij}^2 respectively; b_{0i} , b_{1i} and b_{2i} are the random effects; and ε_{ij} is the measurement error.

6.1 Exploring the covariance structure of the model

In this type of longitudinal study, there are at least three possible components of variability namely: random effects, serial correlation and measurement error (Diggle *et al.*, 1994; Fanta, 2003). Random effects are the effects that arise from the characteristics of the individual children. Therefore, these effects explain the stochastic variation between children. On the other hand, the repeated measurements (*WdVarFEV1*) on successive occasions of the same child are most likely to be serially dependent. Hence, one would not be able to extract as much information from these dependent measurements as one would be able to extract from the same number of independent measurements. That is, serial correlations mask part of the within-child variation in the data. Finally, because the *WdVarFEV1* measurements are determined using an airwatch machine during data collection, it is natural to expect the existence of measurement error.

Figure 6.1 displays the graph of the estimated variance of the residuals versus time for each three asthma status sub-populations. The variance function seems to be relatively stable and hence a constant variance model is plausible.

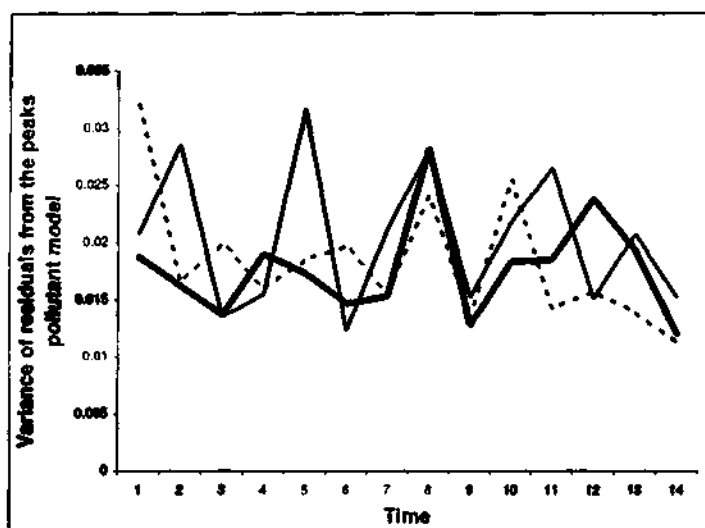


Figure 6.1: Plot of the variance of residuals against time generated from the peaks pollutant model. The solid line represents the persistent asthmatics, the dashed line the asthmatics and the bold line the normal children.

To gain insight into the association among the repeated measurements of $WDVarFEV1$ within-child, a scatter-plot matrix of the residuals is constructed using model (6.1). The scatter plots in Figure 6.2 have circular shapes indicating that the serial correlations are not very strong. Since the effect of serial correlation seems to be dominated by the combination of random effects and the measurement error, the serial correlation is not included in the model.

Conditional on the selected set of random effects, the covariance matrix D for the random effects b_i and the covariance matrix Σ_i for the error components ϵ_i needs to be specified. Although many possible covariance structures are available, Verbeke and Molenberghs (2000) suggest that for longitudinal data, the unstructured covariance structure is the preferred covariance structure for the random effects.

For the covariance structure of the measurement error, a choice has to be made from among the three most commonly used structures in longitudinal mixed models that is: the compound symmetric (CS), autoregressive order one (AR(1)), and the un-

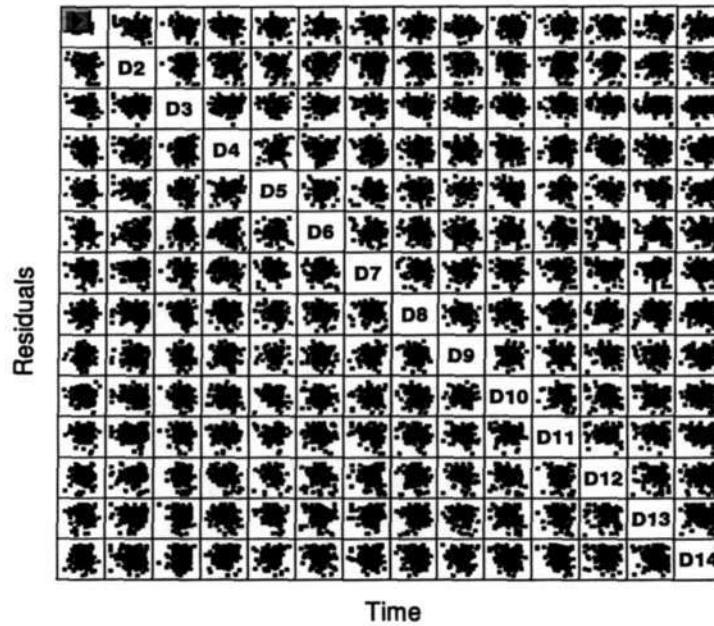


Figure 6.2: Scatter plot of residuals by time. The residuals are generated from the peaks pollutant model. D1 to D14 in the main diagonal represent the 14 repeated measurements.

structured covariance (UN) structures. The application of the unstructured covariance structure to the data set in this study led to the failure of model convergence. Comparison is therefore made only between the compound symmetry and the autoregressive covariance structures. The values of the criteria for the two covariance structures are shown in Table 6.1. Since the covariance structure with the largest values of the criteria is considered to be most desirable, all the three model-fit criteria point to the choice of CS. Therefore, the compound symmetry structure is chosen for the covariance matrix Σ_i . The CS covariance structure will therefore assume the form

$$\Sigma_i = \begin{pmatrix} \sigma^2 + \sigma_c^2 & \sigma_c^2 & \dots & +\sigma_c^2 \\ \sigma_c^2 & \sigma^2 + \sigma_c^2 & \dots & \sigma_c^2 \\ \dots & \dots & \dots & \dots \\ \sigma_c^2 & \sigma_c^2 & \dots & \sigma^2 + \sigma_c^2 \end{pmatrix}.$$

Table 6.1: Covariance structure criteria for the random effects of the peaks pollutant model

Covariance Structure	AIC	AICC	BIC
CS	-2031.1	-2031	-2003.5
AR(1)	-2040.3	-2040.3	-2012.7

6.2 Model reduction

6.2.1 Tests for the significance of the random effects

Using the covariance structure chosen, an investigation is carried out to check the significance of the random effects represented in the model are needed in the model (6.1). In this procedure, all the reduced models obtained by deleting one random effect from model (6.1) are tested. The models and the associated log-likelihood values are shown in Table 6.2. Further more, Table 6.3 shows the LR test statistics for comparing the models.

A mixture of two chi-squared distributions with k_1 and k_2 degrees of freedom, with equal weights 0.5, are therefore denoted by $\chi_{k_1:k_2}^2$. For example, the p -values obtained under the ML estimation for the comparison of Model 2 versus Model 1 can then be calculated as

$$\begin{aligned} p &= P(\chi_{2:3}^2 > 6.7) \\ &= \frac{1}{2}P(\chi_2^2 > 6.7) + \frac{1}{2}P(\chi_3^2 > 6.7). \end{aligned}$$

The naive asymptotic null hypothesis is the one which follows from applying the classical likelihood theory, ignoring the boundary problem for the null hypothesis. In calculation of the p -values, the the covariance structure is simplified by deleting any of the random effects from the model if the computed p -values of the higher random effects are larger than the 5% level of significance.

Table 6.2: The four random-effects models for the peaks pollutant model

Random effects	The Maximum log-likelihood ($\ln(L(\theta))$)	
	ML	REML
Model 1: Intercepts, time, time ²	1232.15	1023.55
Model 2: Intercepts, time	1228.80	1019.80
Model 3: Intercepts	1220.35	1009.50
Model 4: No random effects	1220.35	1009.50

The comparison "Model 4 versus Model 3" has a result of $-2\ln(\lambda_N) = 0$ which means that there is no difference between models 4 and 3. The computations of the p -values are therefore done only for the comparisons "Model 2 versus Model 1" and "Model 3 versus Model 2" and these computed p -values are shown in Table 6.3.

Table 6.3: The likelihood ratio statistic with the correct, the naive asymptotic null distributions for comparing random-effects models and the comparison p -values for the peaks pollutant model

Hypothesis	Likelihood ratio statistic ($-2\ln(\lambda_N)$)		Asymptotic null distribution		P-value	
	ML	REML	Correct	Naive	ML	REML
Model 2 versus Model 1	6.7	7.9	$\chi^2_{2,3}$	χ^2_3	0.05859	0.03369
Model 3 versus Model 2	16.9	20.2	$\chi^2_{1,2}$	χ^2_2	0.00013	0.00002
Model 4 versus Model 3	0	0	$\chi^2_{0,1}$	χ^2_1	—	—

The REML procedure provides p -values that are smaller than 0.05 for all comparisons which means that for this procedure the covariance structure can not be simplified by deleting any of the random effects in model 1. The ML procedure, however, gives p -values that are larger than 0.05 for the comparison "Model 2 versus Model 1", and p -values that are smaller than 0.05 for the comparison "Model 3 versus Model 2". This means that using the ML procedure, the covariance structure can be simplified by deleting the highest random effect which is the squared time effect. However, since the REML estimation performs slightly better than the ML estimation method (Section

4.4.2), the results from the REML estimation are considered. Thus, no random effect is deleted from the model.

6.2.2 Test for the significance of the fixed effects

With the final covariance structure for the model selected, the preliminary mean structure can be improved. The mean structure is improved by including in the model significant interactions of the child-specific variables among themselves. Since there are five child-specific variables, then there are 10 possible 2-way interactions, 10 possible 3-way interactions, 5 possible 4-way interactions, and 1 possible 5-way interaction.

The LR test, using the ML procedure, is therefore used to see whether or not the model with the preliminary mean structure significantly performs better with the introduction of an additional interaction effects of the child-specific variables. As discussed in Section 4.4.1, this test can be used to compare models with different mean structures, and that test is only valid if the models are fitted using the ML procedure.

Let the preliminary model be denoted by M_0 and the model with the additional 2-way interaction terms be denoted by M_1 . To compare the two models, the -2 log-likelihood values for the two models are obtained using the SAS software. Using these values, the LR test statistic $-2\ln(\lambda_N)$ is computed by calculating the difference $(-2l_0) - (-2l_1)$ where l_0 is the log-likelihood of model M_0 and l_1 is the log-likelihood of model M_1 .

Also calculated are the additional degrees of freedom $(m_1 - m_0)$ where m_1 and m_0 are the respective number of parameters in the models M_1 and M_0 . The LR test statistic is then compared with the chi-square value $\chi_{\alpha, m_1 - m_0}^2$. Only those 2-way interactions with p -values less than 0.05 are retained in the model. The LR test statistic, the degrees of freedom and the p -values for testing the significance of the 2-way interactions among

the child-specific factors, 2-way interactions with time, and 2-way interactions with squared time are shown in Table 6.4. The reduced peaks pollutant model (without interactions among the child-specific factors) was found to have $-2l_0 = -2464.3$. The 3-way to the 5-way interaction effects among the child-specific factors were found to be negligible.

Interactions $time * School * AsthmaStatus$ and $time^2 * School * AsthmaStatus$ are found to be significant at a 5% level of significance and are thus added to the peaks pollutant model. In an attempt to reduce the model even further, the backward elimination method is applied.

Table 6.4: The likelihood ratio test statistics and p -values for the 2-way interactions, 2-way interactions with time, and 2-way interactions with squared time in the peaks pollutant model

2-way interactions	Additional degrees of freedom (m1 m0)	2-way interactions		time * 2-way interactions		time ² * 2-way interactions	
		Likelihood ratio test statistic	P-Value	Likelihood ratio test statistic	P-Value	Likelihood ratio test statistic	P-Value
School*Sex	4	3.5	0.4779	4.3	0.3669	4.2	0.3796
School*AsthmaStatus	8	14.5	0.0696	15.7	0.0469	15.6	0.0485
School*Weight	8	11.4	0.1800	2.3	0.9704	2.1	0.9778
School*Height	8	6.4	0.6025	13.0	0.1118	13.9	0.0844
Sex*AsthmaStatus	2	1.3	0.5220	3.6	0.1653	4.6	0.1003
Sex*Weight	2	0.2	0.9048	0.3	0.9807	0.6	0.7408
Sex*Height	2	0.2	0.9048	0.0	1.0000	0.1	0.7518
AsthmaStatus*Weight	4	5.6	0.2311	3.8	0.4337	0.9	0.9246
AsthmaStatus*Height	4	1.9	0.7541	5.8	0.2146	4.9	0.2977
Weight*Height	4	1.6	0.8088	0.4	0.9825	0.4	0.9825

In this application, all the two potential 2-way interactions are introduced in the model and the one with the largest p -value is identified and removed from the model. Since the 2-way interactions $time * School * AsthmaStatus$ and $time^2 * School * AsthmaStatus$ have p -values 0.1232 and 0.1260 respectively, $time^2 * School * AsthmaStatus$ is dropped

and the model fitted again. The p -value for the remaining 2-way interaction, $time * School * AsthmaStatus$, is 0.0469 which is less than 0.05, therefore this 2-way interaction is left in the peaks pollutant model.

6.3 Missingness

The data from the Durban-South health study is characterized by missingness in the response variable WDVArFEV1. Since 55 persistent asthmatics, 63 asthmatics and 115 normal children were considered for analysis, then for each day a total of 55, 63 and 115 responses were expected from these respective sub-populations.

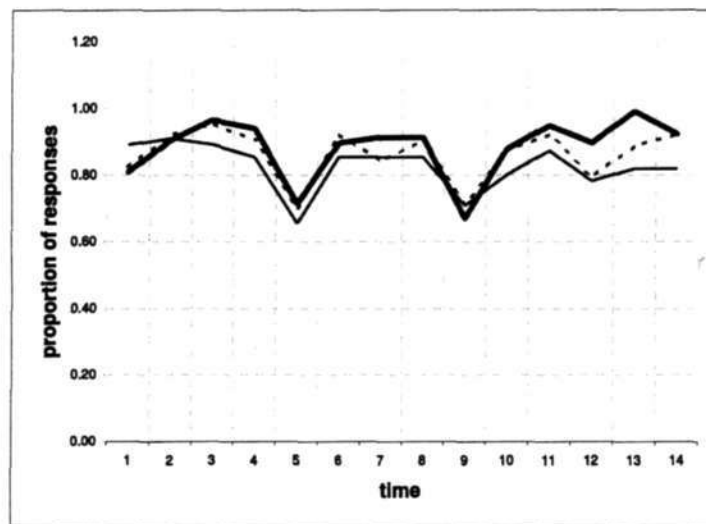


Figure 6.3: Graphical representation of the proportion of respondents to WDVArFEV1 by time. The solid line represents the persistent asthmatics, the dashed line the asthmatics and the bold line the normal children.

However, as shown in the Figure 6.3, not all responses were obtained either due to the FEV1 reading being invalid or due to the child not being at school to participate in the blows on that particular day. Figure 6.3 shows the least responses on time 5 and time 9 which are the dates 04/06/2004 and 10/06/2004 respectively. The low

response on time 5 was due to non participation of the children in Nizam road in the blow exercise. Probably, Nizam road was closed on the particular date. However the low response on time 9 was mainly attributed to the invalid records registered by the children in Ferndale school. This probably could attributed to the malfunctioning of the airwatches or the improper use of the airwatches by the children.

This is not a case of dropout since a child with no response at a particular time may have responses on the times that follow. Therefore, the missing values are intermittent. In order to effectively utilize the PROC MIXED procedure for analysis, the ignorability assumption of missingness has to hold. To test for the type of missingness, an MAR test is applied to the data set. Since the missingness exists for the response variable, then the MAR test for multivariate data cannot be applied since it deals with missingness in the independent variable. In addition, the missingness is intermittent, therefore before the logistic regression model MAR test can be applied, the test has to be modified.

6.3.1 Applying the logistic regression model MAR test

In order to apply the logistic regression model MAR test (Section 4.10.1) to the data set, the following alterations are made to the method. Since the data set is characterized by an intermittent mode of missingness, the missing data process can be explored by assuming that the probability of missingness $P(R_{ij}=0|y_i)$ at time j depends on both the current outcome y_{ij} and the previous one y_{ij-1} . The values of y_{ij} and y_{ij-1} are obtained whenever there is a case of missingness. For example, whenever there is a missing value at y_{ij+1} , the outcomes y_{ij} and y_{ij-1} are included as observations to be applied to the model (4.17) for child i at time j .

Values of the probabilities of the current and previous outcomes are applied to the model (4.17) which is the full model (representing MNAR) and the two reduced

models with $\phi_1 = 0$ and both $\phi_1 = \phi_2 = 0$ (representing MAR and MCAR respectively). These three models are fitted and tested for significant differences using the LR test. In this test, the $-2 \times \log$ -likelihood values for each model are obtained and the difference compared with the chi-square value at 5% level of significance. The degrees of freedom is the difference in the number parameters of the models being compared. A model is said to be significantly different from the other when the attained p -values are less than 0.05. The parameter estimates, log-likelihood values and p -values are shown in Table 6.5.

Table 6.5: The logistic regression model MAR test results for the response variable

Parameter	MCAR	MAR	MNAR
Φ_0	-1.5318	-1.5057	-1.4955
Φ_1	0.0000	0.0000	-0.0434
Φ_2	0.0000	-0.0619	-0.0439
log likelihood	-1654.3067	-1657.8900	-1659.5031
-2log likelihood	3308.6134	3315.3800	3319.0062
Difference with MAR	6.7466		3.6462
P-Values	0.0094		0.0562
Result	Reject MCAR		Reject MNAR
Conclusion		MAR	

The models for MCAR and MAR are significantly different. Therefore the larger model, which is the model representing MAR, is preferable. In comparing the models representing MAR and MNAR, the MAR is found to be preferable to the full model since the two models are found not to be significantly different. It can thus be concluded that the MAR assumption holds. Both MAR tests above show that, conditional to the observed information, the missingness is indeed independent of the unobserved

information. Since the ignorability (specifically MAR) assumption has been found to hold, the PROC MIXED procedure is appropriate to use for the analysis.

6.4 Model checking

To check whether the model fits the data set well, the observed and fitted/predicted WVarFEV1 profiles are compared (Fanta, 2003). In order to do this, the loess smoothing technique is applied to summarize the trend of the observed and predicted WVarFEV1 values as a function of time. The superimposed observed and fitted profiles are shown in Figure 6.4. It can be seen that the profiles are very close together from which it can be concluded that the model fit the data set well.

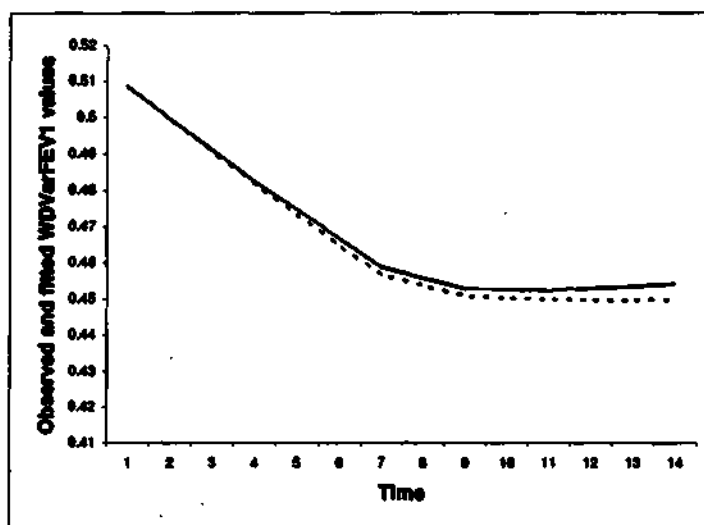


Figure 6.4: The observed and predicted/fitted profiles of the WVarFEV1 values for the peaks pollutant model. The observed profiles are represented by the solid line while the dotted line represents the fitted profiles.

A simultaneous check for normality and outlier detection is also done for the random effects and the residuals. The histogram of the EB estimates and the residuals are shown in Figure 6.5 (a) and (b) respectively. Since both figures show just a symmetric

bell-shaped curve, then the normality assumption is not violated for both the random effects and the residuals

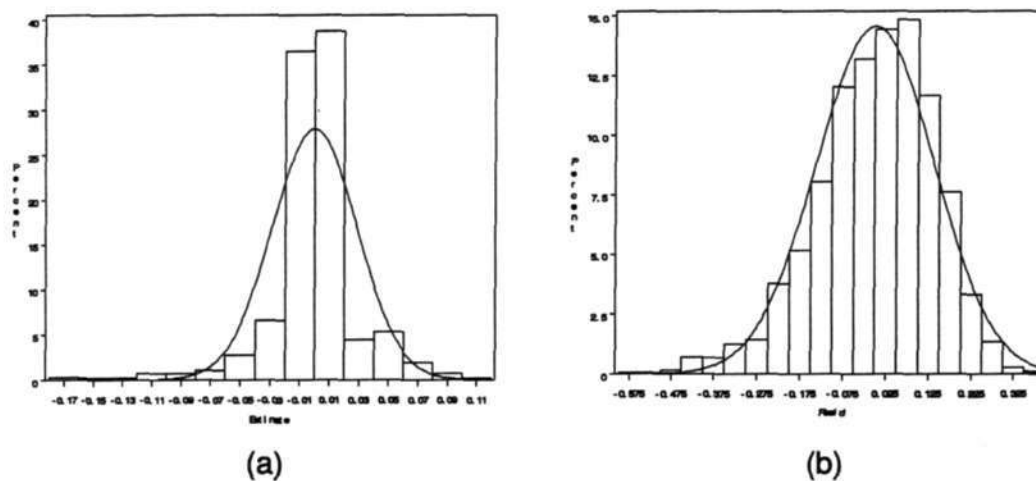


Figure 6.5: Histogram of the (a) EB estimates (for the random effects) and the (b) residuals for the peaks pollutant model.

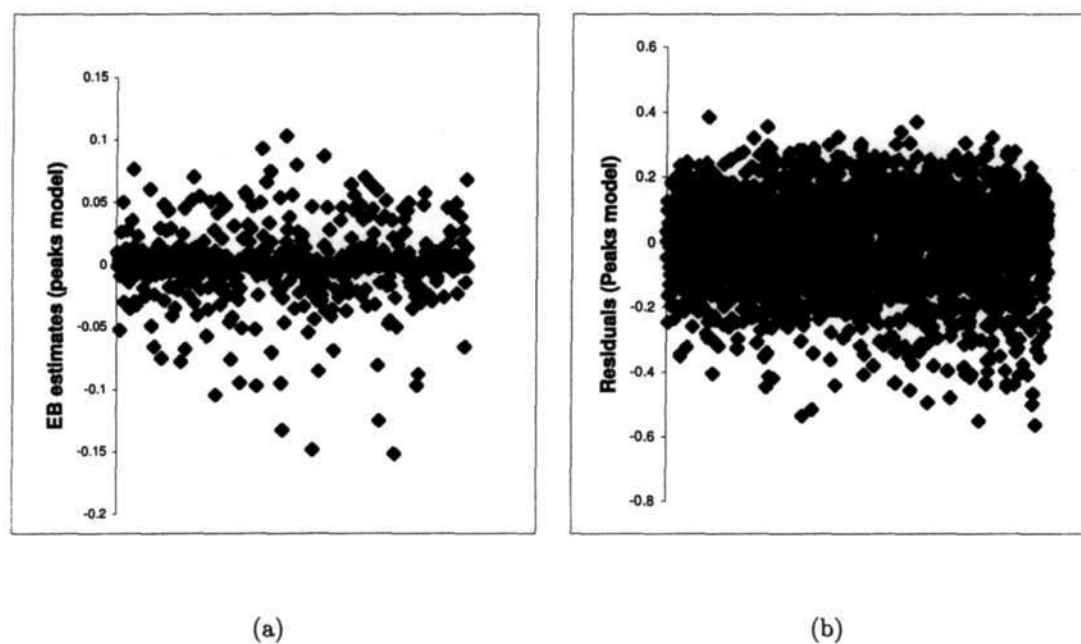


Figure 6.6: Scatter plots of the (a) EB estimates and (b) residuals for the peaks pollutant model.

To identify the presence of outlying observations, the scatter plots of the EB esti-

mates and the residuals, shown in Figure 6.6 are used. The scatter plots show just a few outlying EB estimates and residuals. In addition, the histograms of the EB estimates and residuals (Figure 6.5), which appear just slightly skewed to the left, indicate that there are just a few outliers which may not affect the inference for the model effects.

6.5 Inference from the peaks pollutant model

In this section the estimates of all the model parameters are determined. The REML estimates, standard deviations and p -values are shown in Table 6.6. The aim is to determine which model parameters have a significant effect on the independent variable *WVarFEV1*.

The table shows that all the intercepts in the model are significantly different from zero. More specifically, the intercepts for the children in schools Nizam, Assegai, Dirkie Uys, Ferndale and Ngazana; male and female children; persistent asthmatic, asthmatic and children with no case of asthma; underweight, overweight and children of normal weight; and short, tall and children of normal height, have an intercept significantly greater than zero at a 5% level of significance. The pollutant variable *Peak5* has a significant effect on the independent variable at a 5% level of significance. In other words, a five day prior exposure to a unit increase in the pollutant measure significantly leads to a fall in the *WVarFEV1* measure at a 5% level of significance.

The *WVarFEV1* measure of the following children have a significant linear relationship with time at a 5% level of significance: children in Ngazana; male and female children; children with no cases of asthma; overweight children and children of normal weight; short, tall and children of normal weight; asthmatic children in Ferndale; and children with no case of asthma in Ngazana. In addition the *WVarFEV1* measure of the following children have a significant quadratic coefficient with time: children of

normal weight; short children; and children of normal height.

Table 6.6: The REML estimates, standard deviation and p -values for the peaks pollutant model

Effect	Parameter	Peaks (REML)			
		Estimate	Standard Error	P-value	
Intercepts:					
School	Nizam Road	γ_{01}	0.5138	0.05697	<.0001
	Assegai	γ_{02}	0.5395	0.05483	<.0001
	Dirkie Uys	γ_{03}	0.3794	0.06514	<.0001
	Ferndale	γ_{04}	0.5322	0.05587	<.0001
	Ngazana	γ_{05}	0.5985	0.05431	<.0001
Sex	Male	δ_{01}	0.6652	0.05968	<.0001
	Female	δ_{02}	0.5985	0.05431	<.0001
AsthmaStatus	PAsthmatic	θ_{01}	0.5374	0.05654	<.0001
	Asthmatic	θ_{02}	0.5829	0.05744	<.0001
	Normal	θ_{03}	0.5985	0.05431	<.0001
Weight	Underweight	λ_{01}	0.4443	0.0647	<.0001
	Normal weight	λ_{02}	0.5581	0.05296	<.0001
	Overweight	λ_{03}	0.5985	0.05431	<.0001
Height	Short	α_{01}	0.7072	0.05099	<.0001
	Normal Height	α_{02}	0.6487	0.03836	<.0001
	Tall	α_{03}	0.5985	0.05431	<.0001
Pollutant:					
	Peak1	β_1	0.001983	0.001351	0.1423
	Peak2	β_2	0.001571	0.002764	0.5697
	Peak3	β_3	-0.00076	0.001264	0.5465
	Peak4	β_4	0.000114	0.001086	0.9165
	Peak5	β_5	-0.0047	0.00122	0.0001
Time Effects:					
School	Nizam Road	γ_{11}	-0.01414	0.01701	0.4070
	Assegai	γ_{12}	-0.0243	0.01624	0.1361
	Dirkie Uys	γ_{13}	0.007183	0.01948	0.7127
	Ferndale	γ_{14}	-0.03158	0.0167	0.0600
	Ngazana	γ_{15}	-0.03462	0.01622	0.0341
Sex	Male	δ_{11}	-0.04026	0.01779	0.0247
	Female	δ_{12}	-0.03462	0.01622	0.0341
AsthmaStatus	PAsthmatic	θ_{11}	-0.02581	0.01706	0.1318
	Asthmatic	θ_{12}	-0.03163	0.0172	0.0675
	Normal	θ_{13}	-0.03462	0.01622	0.0341
Weight	Underweight	λ_{11}	-0.02535	0.01918	0.1878
	Normal weight	λ_{12}	-0.03382	0.01575	0.0331
	Overweight	λ_{13}	-0.03462	0.01622	0.0341
Height	Short	α_{11}	-0.03976	0.01506	0.0089
	Normal Height	α_{12}	-0.03428	0.01144	0.0031
	Tall	α_{13}	-0.03462	0.01622	0.0341
School*Status	Nizam*PAsthma	γ_{111}	-0.00383	0.01764	0.8284
	Nizam*Asthma	γ_{112}	-0.00911	0.0179	0.6112
	Nizam*Normal	γ_{113}	-0.01414	0.01701	0.4070
	PAsthm*Assegai	γ_{121}	-0.01161	0.01633	0.4780
	Asthma*Assegai	γ_{122}	-0.02701	0.01728	0.1196
	Normal*Assegai	γ_{123}	-0.0243	0.01624	0.1361
	PAsthma*Dirkie	γ_{131}	0.0185	0.01999	0.3560
	Asthma*Dirkie	γ_{132}	-0.00127	0.01953	0.9484
	Normal*Dirkie	γ_{133}	0.007183	0.01948	0.7127
	PAsthm*Ferndale	γ_{141}	-0.01696	0.01759	0.3361
	Asthma*Ferndale	γ_{142}	-0.03711	0.01761	0.0364
	Normal*Ferndale	γ_{143}	-0.03158	0.0167	0.0600
	PAsthm*Ngazana	γ_{151}	-0.02581	0.01706	0.1318
	Asthma*Ngazana	γ_{152}	-0.03163	0.0172	0.0675
	Normal*Ngazana	γ_{153}	-0.03462	0.01622	0.0341

Table 6.6: The REML estimates, standard deviation and p -values for the peaks pollutant model
(Continued)

Effect	Parameter	Peaks (REML)			
		Estimate	Standard Error	P-value	
Time² Effects:					
School	Nizam Road	γ_{21}	0.000764	0.001099	0.4878
	Aasegai	γ_{22}	0.001396	0.001041	0.1844
	Dirkie Uys	γ_{23}	-0.000510	0.001235	0.6809
	Ferndale	γ_{24}	0.001783	0.001073	0.0683
	Ngazana	γ_{25}	0.002028	0.001047	0.0541
Sex	Male	δ_{21}	0.002100	0.001146	0.0685
	Female	δ_{22}	0.002028	0.001047	0.0541
AsthmaStatus	PAsthmatic	θ_{21}	0.001400	0.001088	0.1998
	Asthmatic	θ_{22}	0.002101	0.001106	0.0589
	Normal	θ_{23}	0.002028	0.001047	0.0541
Weight	Underweight	λ_{21}	0.001751	0.001235	0.1580
	Normal weight	λ_{22}	0.002014	0.001014	0.0484
	Overweight	λ_{23}	0.002028	0.001047	0.0541
Height	Short	α_{21}	0.002120	0.000966	0.0294
	Normal Height	α_{22}	0.002016	0.000734	0.0066
	Tall	α_{23}	0.002028	0.001047	0.0541
Covariance of b_i:					
	$\text{Var}(b_{0i})$	d_{11}	0.005974	0.002724	0.0141
	$\text{Cov}(b_{0i}, b_{1i})$	$d_{12}=d_{21}$	-0.000340	0.000708	0.6329
	$\text{Var}(b_{1i})$	d_{22}	0.000380	0.000235	0.0530
	$\text{Cov}(b_{0i}, b_{2i})$	$d_{13}=d_{31}$	0.000006	0.000043	0.8540
	$\text{Cov}(b_{1i}, b_{2i})$	$d_{23}=d_{32}$	-0.000020	0.000015	0.1345
	$\text{Var}(b_{2i})$	d_{33}	0.000002	0.000000	.
Residual variance:					
	$\text{Var}(e_{ij})$	σ^2	0.000453	0.000014	<.0001
	CS	σ_c^2	0.020610	0.000633	<.0001

To determine whether the mean effect of the Durban-South schools is equal to that of the schools located in the northern residential areas, the hypothesis

$$H_{01} : \frac{\gamma_{01} + \gamma_{02} + \gamma_{03}}{3} = \frac{\gamma_{04} + \gamma_{05}}{2}$$

$$H_{A1} : \frac{\gamma_{01} + \gamma_{02} + \gamma_{03}}{3} \neq \frac{\gamma_{04} + \gamma_{05}}{2}$$

is tested. The test results in an F -value of 8.33 and p -value of 0.0003. Since the p -value is less than 0.05, the null hypothesis H_{01} is rejected in favour of H_{A1} . Therefore, at a 5% level of significance, the mean effects of the schools in the two regions are not equal.

The hypothesis

$$H_{02} : \delta_{01} = \delta_{02}$$

$$H_{A2} : \delta_{01} \neq \delta_{02}$$

can be used to test whether the mean effects of the male and female children are equal. The resulting F -value and p -value are 8.25 and 0.0045 respectively. Since the p -value is less than 0.05, the null hypothesis H_{02} is rejected in favour of H_{A2} . We therefore conclude that, at a 5% level of significance, the mean effects of the male and female children are not equal.

To test if the mean effects of the three asthma status categories are equal, the hypothesis

$$H_{03} : \theta_{01} = \theta_{02} = \theta_{03}$$

$$H_{A3} : \text{At least two of the asthma status means are different}$$

is tested. The results show an F -value of 2.44 and p -value of 0.0898. Since the p -value is greater than 0.05, we fail to reject the null hypothesis, H_{03} . The data therefore suggests that the mean effects of the three asthma status categories are equal.

To compare if the mean effects of the three weight categories are equal, the hypothesis

$$H_{04} : \lambda_{01} = \lambda_{02} = \lambda_{03}$$

$$H_{A4} : \text{At least two of the mean effects are different}$$

is tested. The results provide an F -value of 5.98 and a p -value of 0.0030. Since the null hypothesis H_{04} is rejected in favour of H_{A4} , a further investigation is carried out

to determine which mean effects are significantly different. The hypotheses tested are:

$$H_{05} : \lambda_{01} = \lambda_{02}$$

$$H_{A5} : \lambda_{01} \neq \lambda_{02},$$

$$H_{06} : \lambda_{01} = \lambda_{03}$$

$$H_{A6} : \lambda_{01} \neq \lambda_{03}, \text{ and}$$

$$H_{07} : \lambda_{02} = \lambda_{03}$$

$$H_{A7} : \lambda_{02} \neq \lambda_{03}.$$

The hypothesis H_{05} versus H_{A5} tests whether the mean effects of the underweight children and children of normal weight are equal. The F -value of this test is found to be 9.49 while the p -value is found to be 0.0024. Since the p -value is very small, it can be concluded that the mean effects of the underweight and children of normal weight are not equal.

The hypothesis H_{06} versus H_{A6} tests whether the mean effects of the underweight and overweight children are equal. The F -value and p -value of this test is found to be 11.02 and 0.0011 respectively. From this it can be concluded that the mean effects of the underweight and overweight children are not equal. Finally, the hypothesis H_{07} versus H_{A7} tests whether the mean effects of the overweight children and children of normal weight are equal. The F -value is found to be 1.6 while the p -value is found to be 0.2016. Since the p -value is large, we fail to reject the null hypothesis H_{07} and conclude that the overweight and children of normal weight have equal mean effects.

A comparison of the three height categories can be done by testing the hypothesis

$$H_{08} : \alpha_{01} = \alpha_{02} = \alpha_{03}$$

$H_{A8} :$ At least two of the mean effects are different.

The results provide an F -value of 1.90 and a p -value of 0.1529. Since the p -value is greater than 0.05, we fail to reject the null hypothesis H_{08} concluding that the mean effects of the height categories are equal at a 5% level of significance.

To compare the linear slopes of the male and female children, the hypothesis

$$H_{09} : \delta_{11} = \delta_{12}$$

$$H_{A9} : \delta_{11} \neq \delta_{12}$$

is tested. The results provide an F -value of 0.68 and p -value of 0.4117 from which it can be concluded that the linear slopes for the male and female children are equal at a 5% level of significance. The comparison of the linear slopes for the three height categories is done by testing the hypothesis

$$H_{010} : \alpha_{11} = \alpha_{12} = \alpha_{13}$$

$H_{A10} :$ At least two of the slopes are different.

This results in an F -value of 0.13 and p -value of 0.8740 from which it can be concluded that the linear slopes for three height categories are equal at a 5% level of significance.

To compare whether the linear slope for overweight children is equal to that for children of normal weight, the hypothesis

$$H_{011} : \lambda_{12} = \lambda_{13}$$

$$H_{A11} : \lambda_{12} \neq \lambda_{13}$$

is tested. The resulting F -value and p -value are found to be 0.01 and 0.9330 respectively. We fail to reject the null hypothesis and conclude that the linear slopes are equal.

For the quadratic coefficients, the coefficients of the short children and children of normal height are compared by testing the hypothesis

$$H_{011} : \alpha_{21} = \alpha_{22}$$

$$H_{A11} : \alpha_{21} \neq \alpha_{22}$$

This results in an F -value of 0.02 and a p -value of 0.8780. Since the p -value is large, we fail to reject the null hypothesis and conclude that the quadratic coefficients are not equal.

To compare the WVarFEV1 values of the children in the Durban-South schools and those in the schools located in the northern residential areas, a graphical representation of superimposed fitted WVarFEV1 values for children in the two regions is used. To do this, the loess smoothing technique is applied to summarize the trend of the fitted WVarFEV1 values as a function of time.

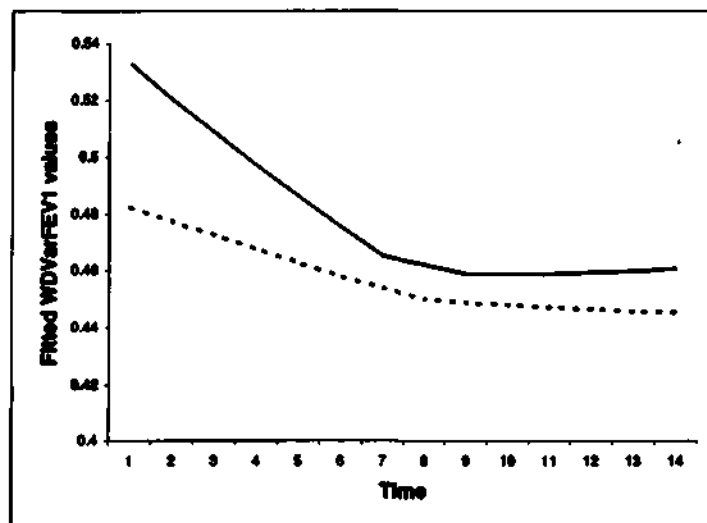


Figure 6.7: The peaks pollutant model fitted profiles of the WVarFEV1 values for children in Durban-South schools and those located in the northern residential areas. The solid line represents the Durban-South profiles while the dotted line represents the northern residential area profiles.

The superimposed graph shown in Figure 6.7 shows that the WVarFEV1 values of

the children in the Durban-South schools are on average greater than the WDV_{Var}FEV₁ values of the children in the schools located in the northern residential areas

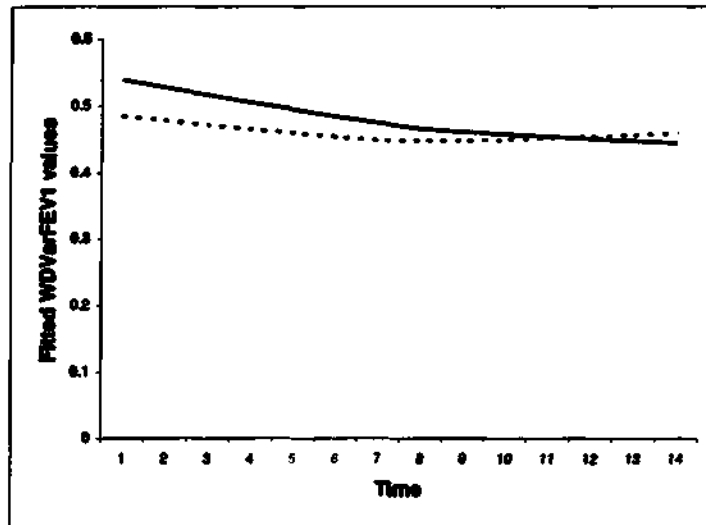


Figure 6.8: The peaks pollutant model fitted profiles of the WDV_{Var}FEV₁ values for the male and female children. The solid line represents the profile for the male children while the dotted line represents the profile for the female children.

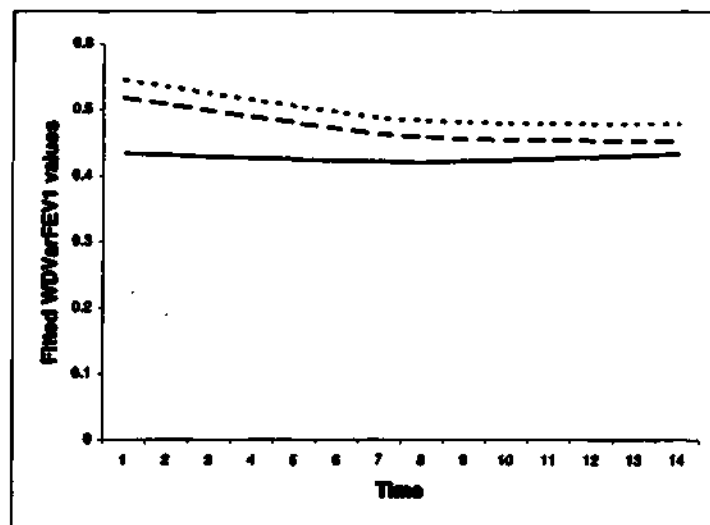


Figure 6.9: The peaks pollutant model fitted profiles of the WDV_{Var}FEV₁ values for children of the three weight categories. The solid line, dashed line and dotted line represent the profiles of the underweight children, children of normal weight and overweight children respectively.

Similarly, Figure 6.8 compares the profiles of the male and female children. It can be seen from this figure that indeed the males and female children do not have a common slope or intercept. Figure 6.9, which compares the profiles of the three weight categories, shows that the overweight children have the highest WDV_{FEV1} values followed by the WDV_{FEV1} values of the children of normal weight.

Chapter 7

Analysis of the 8-hour maximum pollutant model

In this chapter, the 8-hour maximum model is discussed. First investigated is the covariance structure of the model

$$\begin{aligned} y_{ij} = & \sum_{c=1}^5 \gamma_{0c} \text{School}_{ci} + \sum_{s=1}^2 \delta_{0s} \text{Sex}_{si} \\ & + \sum_{u=1}^3 \theta_{0u} \text{AsthmaStatus}_{ui} + \sum_{w=1}^3 \lambda_{0w} \text{Weight}_{wi} + \sum_{h=1}^3 \alpha_{0h} \text{Height}_{hi} \\ & + \sum_{c=1}^5 \gamma_{1c} \text{School}_{ci} t_{ij} + \sum_{s=1}^2 \delta_{1s} \text{Sex}_{si} t_{ij} \\ & + \sum_{u=1}^3 \theta_{1u} \text{AsthmaStatus}_{ui} t_{ij} + \sum_{w=1}^3 \lambda_{1w} \text{Weight}_{wi} t_{ij} + \sum_{h=1}^3 \alpha_{1h} \text{Height}_{hi} t_{ij} \\ & + \sum_{c=1}^5 \gamma_{2c} \text{School}_{ci} t_{ij}^2 + \sum_{s=1}^2 \delta_{2s} \text{Sex}_{si} t_{ij}^2 \\ & + \sum_{u=1}^3 \theta_{2u} \text{AsthmaStatus}_{ui} t_{ij}^2 + \sum_{w=1}^3 \lambda_{2w} \text{Weight}_{wi} t_{ij}^2 + \sum_{h=1}^3 \alpha_{2h} \text{Height}_{hi} t_{ij}^2 \\ & + \beta_1 \text{HrMax1}_{ij} + \beta_2 \text{HrMax2}_{ij} + \beta_3 \text{HrMax3}_{ij} + \beta_4 \text{HrMax4}_{ij} + \beta_5 \text{HrMax5}_{ij} \\ & + b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \varepsilon_{ij}, \end{aligned}$$

where the variables are the same as described in (6.1) except that the five pollutant variables in this case are the 8-hour maximum variables $HrMax1, \dots, HrMax5$. A relatively stable variance function shown in Figure 7.1 shows that a constant variance is plausible for the model. The scatter-plot matrix of residuals in Figure 7.2 has a circular shape indicating that the serial correlation is not very strong. Therefore the serial correlation is not included in the model.

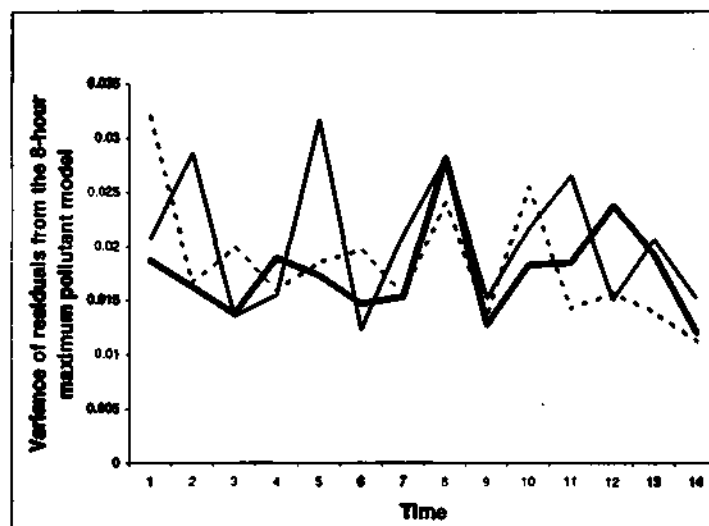


Figure 7.1: Plot of the variance of residuals against time generated from the 8-hour maximum pollutant model. The solid line represents the persistent asthmatics, the dashed line the asthmatics and the bold line the normal children.

For the random effects, the unstructured covariance structure is the preferred covariance structure. However, for the residual, all the three model-fit criteria shown in Table 7.1 point to the CS. Therefore, the compound symmetry structure is chosen for the covariance matrix Σ_{ϵ} in the model.

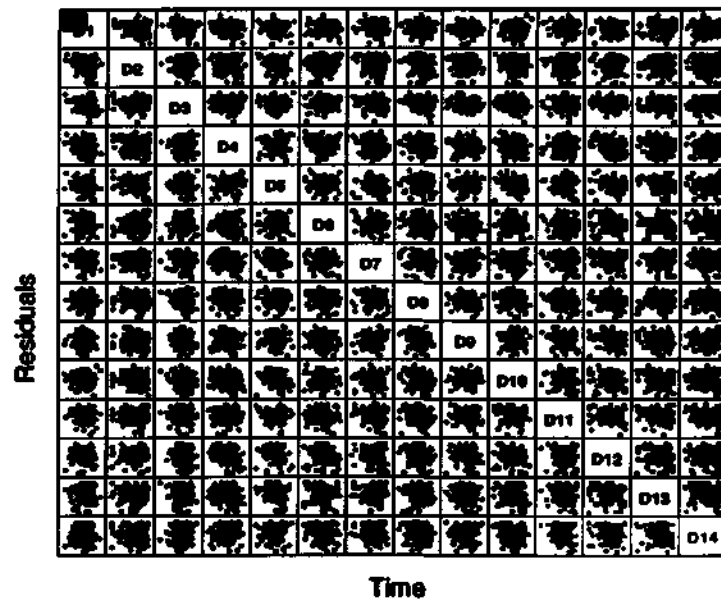


Figure 7.2: Scatter plot of residuals by time. The residuals are generated from the 8-hour maximum pollutant model. D1 to D14 in the main diagonal represent the 14 repeated measurements.

Table 7.1: Covariance structure criteria for the random effects of the 8-hour maximum pollutant model

Covariance Structure	AIC	AICC	BIC
CS	-2010.9	-2010.9	-1983.3
AR(1)	-2020.6	-2020.6	-1993

7.1 Model reduction of the model

The models and the associated maximized log-likelihood values for the 8-hour maximum model are shown in Table 7.2, while Table 7.3 shows the LR statistics for dropping one random effect at a time, starting from the quadratic time effect. These tables show no difference between models 4 and 3. The computations of the p -values

are done for the comparisons “Model 2 versus Model 1” and “Model 3 Versus Model 2” and the computed p -values are shown in Table 7.3.

Table 7.2: The four random-effects models for the 8-hour maximum pollutant model

Random effects	The Maximum log-likelihood ($\ln(L(\theta))$)	
	ML	REML
Model 1: Intercepts, time, time ²	1234.45	1013.45
Model 2: Intercepts, time	1231.10	1009.50
Model 3: Intercepts	1222.65	999.35
Model 4: No random effects	1222.65	999.35

Table 7.3: The likelihood ratio statistic with the correct, the naive asymptotic null distributions for comparing random-effects models and the comparison p -values of the 8-hour maximum pollutant model

Hypothesis	Likelihood ratio statistic ($-2\ln(\lambda_N)$)		P-value	
	ML	REML	ML	REML
Model 2 versus Model 1	6.7	7.9	0.05859	0.03369
Model 3 versus Model 2	16.9	20.3	0.00013	0.00002
Model 4 versus Model 3	0	0		

All estimations in Table 7.3 provide p -values smaller than 0.05 except for the p -values associated with the ML estimation for the comparison “Model 2 versus Model 1”. Since the REML estimation performs slightly better than the ML estimation method, no random effect is deleted from the 8-hour maximum pollutant model.

Table 7.4: The likelihood ratio test statistics and *p*-values for the 2-way interactions, 2-way interactions with time, and 2-way interactions with squared time in the 8-hour maximum pollutant model

2-way interactions	Additional degrees of freedom (m1 m0)	2-way interactions		time * 2-way interactions		time ² * 2-way interactions	
		Likelihood ratio test statistic	P-Value	Likelihood ratio test statistic	P-Value	Likelihood ratio test statistic	P-Value
School*Sex	4	3.5	0.4779	4.2	0.3796	4.1	0.3926
School*AsthmaStatus	8	14.5	0.0696	15.6	0.0485	15.5	0.0501
School*Weight	8	11.2	0.1906	2.2	0.9743	2.0	0.9810
School*Height	8	6.4	0.6025	12.9	0.1153	13.8	0.0671
Sex*AsthmaStatus	2	1.2	0.5488	3.6	0.1653	4.5	0.1054
Sex*Weight	2	0.2	0.9048	0.3	0.8607	0.5	0.7788
Sex*Height	2	0.1	0.9512	0.0	1.0000	0.1	0.7518
AsthmaStatus*Weight	4	5.6	0.2311	3.9	0.4337	0.8	0.9384
AsthmaStatus*Height	4	1.8	0.7725	5.7	0.2227	4.9	0.2977
Weight*Height	4	1.5	0.8266	0.3	0.9696	0.3	0.9696

For the model reduction of the mean structure, the reduced 8-hour maximum pollutant model has $-2l_0 = -2468.9$. The only interaction found to be significant is the 2-way interaction with time $time*School*AsthmaStatus$, with a *p*-value = 0.0485. Therefore only $time*School*AsthmaStatus$ is added to the 8-hour maximum pollutant model.

7.2 Inference of the 8-hour maximum model

Before the inference of the model is done, the model fit is checked by fitting observed and fitted WdVarFEV1 profiles using the loess smoothing technique. The summarized trends of observed and predicted WdVarFEV1 values as a function of time are shown in Figure 7.3. Since the profiles are very close together, it can be concluded that the model fit the data set well.

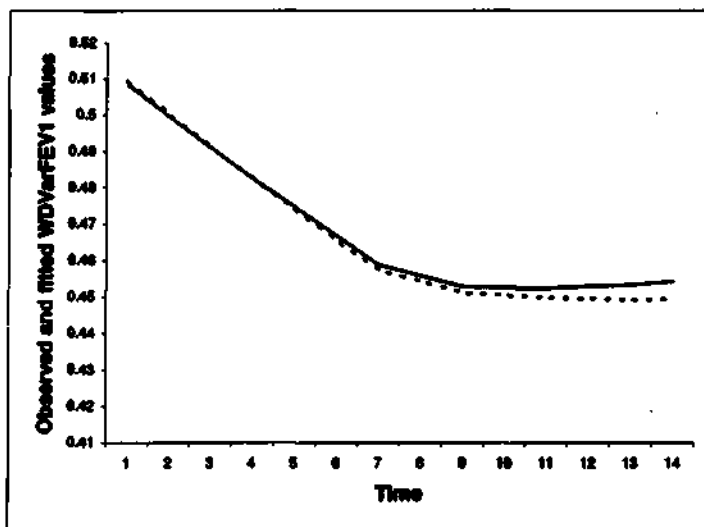


Figure 7.3: The 8-hour maximum pollutant model observed and predicted/fitted profiles of the WDVArFEV1 values. The observed profiles are represented by the solid line while the dotted line represents the fitted profiles.

In addition, the assumption of normality of the random effects and the residual is tested. The histogram of the EB estimates and the residuals are shown in Figure 7.4 (a) and (b) respectively. Both figures show a symmetric bell-shaped curve indicating that the normality assumption is not violated for both the random effects and the residuals. The scatter plots in Figure 7.5 for both the EB and residuals show just a few outliers which may not affect the inference of the model effects. The same conclusion can be drawn from the slight left skewness of the histograms in Figure 7.4.

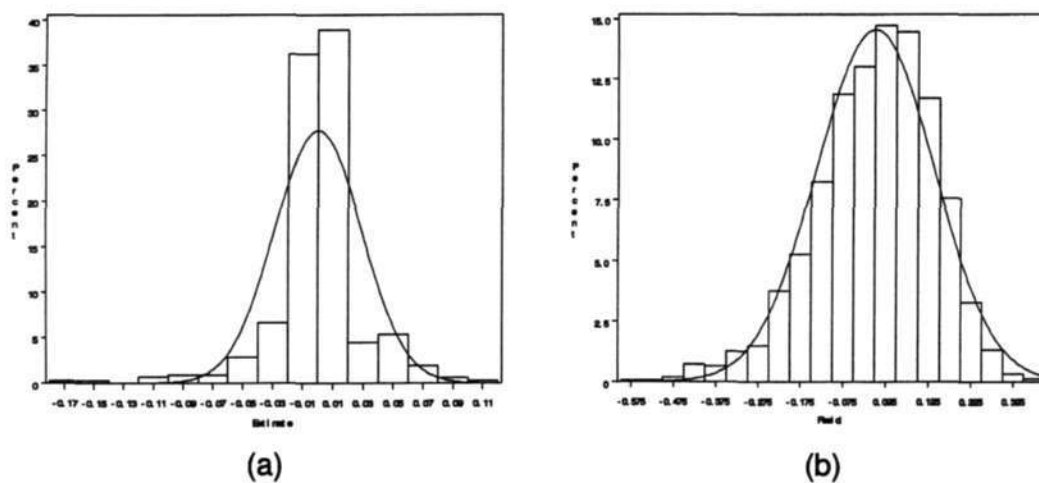


Figure 7.4: Histogram of the (a) EB estimates (for the random effects) and the (b) residuals for the 8-hour maximum pollutant model.

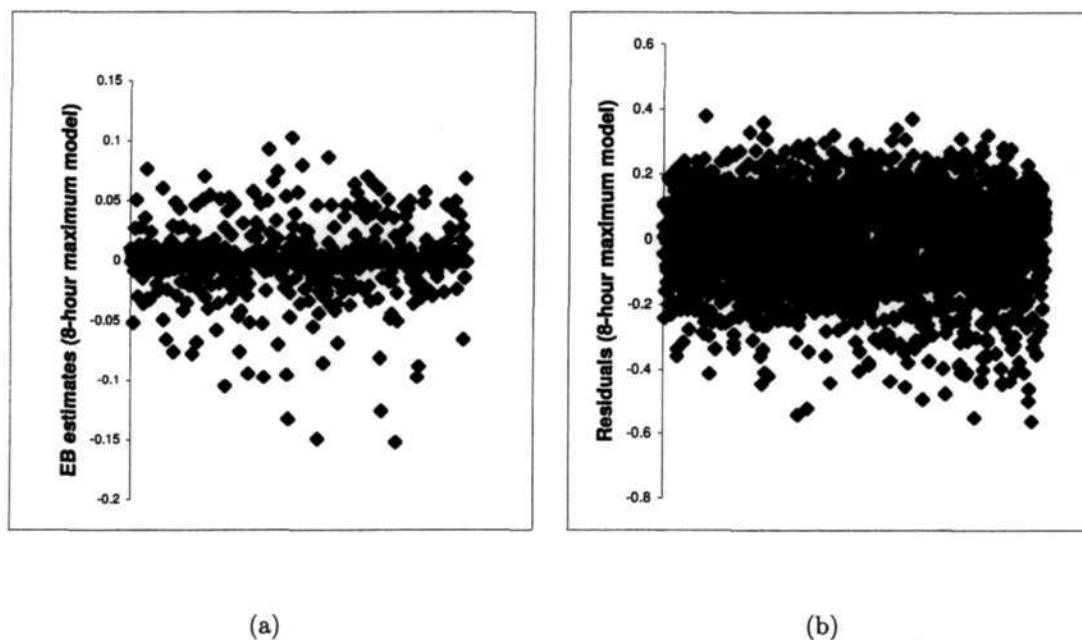


Figure 7.5: Scatter plot of the (a) EB estimates and (b) residuals for the 8-hour maximum pollutant model.

Table 7.5: The REML estimates, standard deviation and p -values for the 8-Hour maximum pollutant model

Effect	Parameter	8-Hour Maximum (REML)			
		Estimate	Standard Error	P-value	
Intercepts:					
School	Nizam Road	γ_{01}	0.531500	0.057650	<.0001
	Assegai	γ_{02}	0.581400	0.058940	<.0001
	Dirkie Uys	γ_{03}	0.401600	0.065550	<.0001
	Ferndale	γ_{04}	0.533000	0.055900	<.0001
	Ngazana	γ_{05}	0.602100	0.054330	<.0001
Sex	Male	δ_{01}	0.669000	0.059710	<.0001
	Female	δ_{02}	0.602100	0.054330	<.0001
AsthmaStatus	PAsthmatic	θ_{01}	0.540700	0.056570	<.0001
	Asthmatic	θ_{02}	0.585800	0.057460	<.0001
	Normal	θ_{03}	0.602100	0.054330	<.0001
Weight	Underweight	λ_{01}	0.447700	0.064740	<.0001
	Normal weight	λ_{02}	0.561000	0.052980	<.0001
	Overweight	λ_{03}	0.602100	0.054330	<.0001
Height	Short	α_{01}	0.710600	0.051010	<.0001
	Normal Height	α_{02}	0.652600	0.038360	<.0001
	Tall	α_{03}	0.602100	0.054330	<.0001
Pollutant:					
	HrMax1	β_1	0.000184	0.000115	0.1100
	HrMax2	β_2	0.000130	0.000196	0.5082
	HrMax3	β_3	-0.000220	0.000113	0.0489
	HrMax4	β_4	-0.000020	0.000097	0.8533
	HrMax5	β_5	-0.000470	0.000112	<.0001
Time Effects:					
School	Nizam Road	γ_{11}	-0.015020	0.017060	0.3797
	Assegai	γ_{12}	-0.028430	0.016180	0.0805
	Dirkie Uys	γ_{13}	0.003781	0.019510	0.8466
	Ferndale	γ_{14}	-0.031100	0.016720	0.0643
	Ngazana	γ_{15}	-0.035030	0.016230	0.0322
Sex	Male	δ_{11}	-0.040760	0.017810	0.0232
	Female	δ_{12}	-0.035030	0.016230	0.0322
AsthmaStatus	PAsthmatic	θ_{11}	-0.026130	0.017070	0.1275
	Asthmatic	θ_{12}	-0.031880	0.017220	0.0656
	Normal	θ_{13}	-0.035030	0.016230	0.0322
Weight	Underweight	λ_{11}	-0.025630	0.019190	0.1833
	Normal weight	λ_{12}	-0.034050	0.015770	0.0320
	Overweight	λ_{13}	-0.035030	0.016230	0.0322
Height	Short	α_{11}	-0.040190	0.015080	0.0083
	Normal Height	α_{12}	-0.034740	0.011460	0.0027
	Tall	α_{13}	-0.035030	0.016230	0.0322
School*Status	Nizam*PAsthma	γ_{111}	-0.004570	0.017680	0.7961
	Nizam*Asthma	γ_{112}	-0.009800	0.017940	0.5856
	Nizam*Normal	γ_{113}	-0.015020	0.017060	0.3797
	PAsthm*Assegai	γ_{121}	-0.015650	0.016270	0.3372
	Asthma*Assegai	γ_{122}	-0.030970	0.017230	0.0738
	Normal*Assegai	γ_{123}	-0.028430	0.016180	0.0805
	PAsthma*Dirkie	γ_{131}	0.015250	0.020030	0.4474
	Asthma*Dirkie	γ_{132}	-0.004400	0.019560	0.8221
	Normal*Dirkie	γ_{133}	0.003781	0.019510	0.8466
	PAsthm*Ferndale	γ_{141}	-0.016390	0.017620	0.3532
	Asthma*Ferndale	γ_{142}	-0.036480	0.017640	0.0399
	Normal*Ferndale	γ_{143}	-0.031100	0.016720	0.0643
	PAsthm*Ngazana	γ_{151}	-0.026130	0.017070	0.1275
	Asthma*Ngazana	γ_{152}	-0.031880	0.017220	0.0656
	Normal*Ngazana	γ_{153}	-0.035030	0.016230	0.0322

Table 7.5 continued

Effect	Parameter	8-Hour Maximum (REML)			
		Estimate	Standard Error	P-value	
Time² Effects:					
School	Nizam Road	γ_{21}	0.000844	0.001102	0.4449
	Assegai	γ_{22}	0.001539	0.001037	0.1393
	Dirkie Uys	γ_{23}	-0.000330	0.001237	0.7898
	Ferndale	γ_{24}	0.001765	0.001075	0.1021
	Ngazana	γ_{25}	0.002058	0.001048	0.0509
Sex	Male	δ_{21}	0.002137	0.001147	0.0640
	Female	δ_{22}	0.002058	0.001048	0.0509
AsthmaStatus	PAsthmatic	θ_{21}	0.001423	0.001089	0.1927
	Asthmatic	θ_{22}	0.002121	0.001107	0.0568
	Normal	θ_{23}	0.002058	0.001048	0.0509
Weight	Underweight	λ_{21}	0.001769	0.001236	0.1539
	Normal weight	λ_{22}	0.002032	0.001014	0.0466
	Overweight	λ_{23}	0.002058	0.001048	0.0509
Height	Short	α_{21}	0.002153	0.000967	0.0271
	Normal Height	α_{22}	0.002047	0.000735	0.0059
	Tall	α_{23}	0.002058	0.001048	0.0509
Covariance of b_j:					
	Var(b_{0j})	d_{11}	0.006040	0.002730	0.0135
	Cov(b_{0j}, b_{1j})	$d_{12}=d_{21}$	-0.000360	0.000709	0.6149
	Var(b_{1j})	d_{22}	0.000386	0.000235	0.0506
	Cov(b_{0j}, b_{2j})	$d_{13}=d_{31}$	0.000009	0.000043	0.8328
	Cov(b_{1j}, b_{2j})	$d_{23}=d_{32}$	-0.000020	0.000015	0.1284
	Var(b_{2j})	d_{33}	0.000002	0.000000	.
Residual variance:					
	Var(ϵ_{ij})	σ^2	0.000469	0.000014	<.0001
	CS	σ_C^2	0.020570	0.000632	<.0001

Table 7.5 aims at investigating which factors in model have an effect on child respiratory conditions, represented by the independent variable $WDVarFEV1$. It can be seen that all the factor levels have significant mean effects, at a 5% level of significance. In addition, for the children in Ngazana; male and female children; children with no case of asthma; children of normal weight and overweight children; short, tall and children of normal height; asthmatic children in Ferndale; and children with no case of asthma in Ngazana, it is found that $WDVarFEV1$ has a linear relationship with time. For children of normal weight; short children; and children of normal height, there is a significant quadratic relationship between $WDVarFEV1$ and time. The pollutant variables $HrMax3$ and $HrMax5$ have a significant effect on the dependent variable. In

other words, three day and five day prior exposure to a unit increase in the pollutant measure significantly results in a fall in the WDV_{Var}FEV1 measure at a 5% level of significance, keeping all the other variables fixed.

The effects of the Durban-South and northern residential area schools (F -value 4.85, p -value 0.0086); child sex (F -value 8.81, p -value 0.0033); and child weight (F -value 6.33, p -value 0.0022) on WDV_{Var}FEV1 are also found to be significant. The equality of the linear slopes for male and female children (F -value 0.74, p -value 0.3911); the height categories (F -value 0.14, p -value 0.8737); and the overweight and children of normal weight (F -value 0.01, p -value 0.9179) are found to be tenable at a 5% level of significance. In addition, the quadratic coefficients of the short children and children of normal weight (F -value 0.02, p -value 0.8761) are also found to be equal at a 5% level of significance.

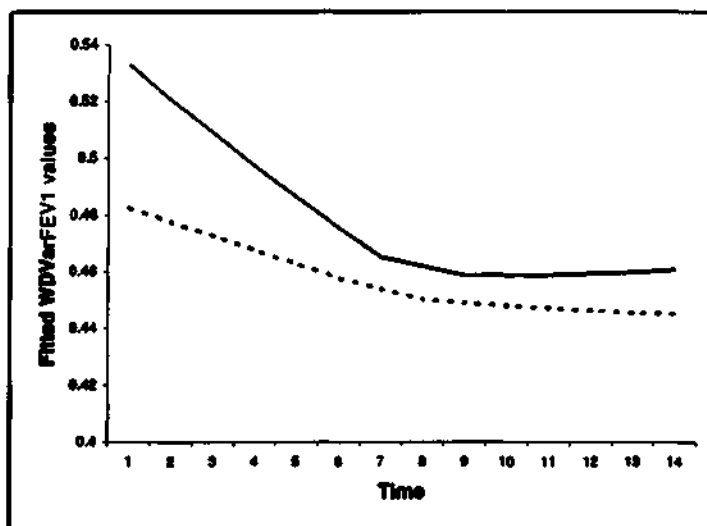


Figure 7.6: The 8-hour maximum pollutant model fitted profiles of the WDV_{Var}FEV1 values for children in Durban-South schools and those located in the northern residential areas. The solid line represents the Durban-South profiles while the dotted line represents the northern residential area profiles.

Using the loess technique, a graphical comparison of the fitted WDV_{Var}FEV1 values

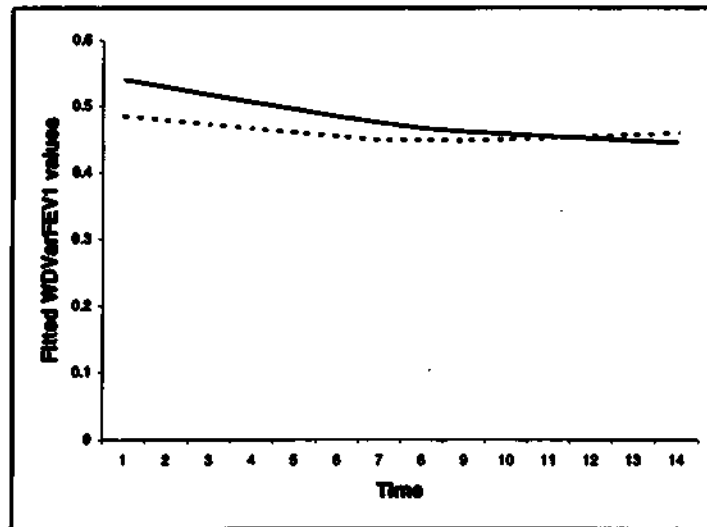


Figure 7.7: The 8-hour maximum pollutant model fitted profiles of the WDVArFEV1 values for the male and female children. The solid line represents the profile for the male children while the dotted line represents the profile for the female children.

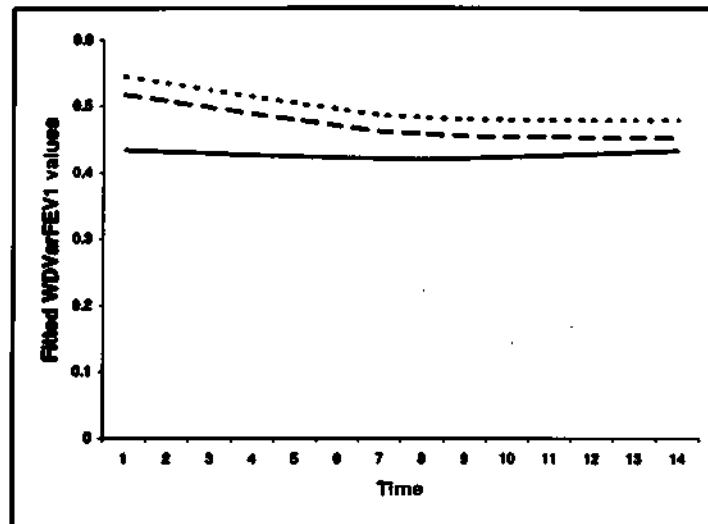


Figure 7.8: The 8-hour maximum pollutant model fitted profiles of the WDVArFEV1 values for children of the three weight categories. The solid line, dashed line and dotted line represent the profiles of the underweight children, children of normal weight and overweight children respectively.

of children in the Durban-South schools and those in the schools located in the northern residential areas (Figure 7.6); male and female children (Figure 7.7); and three weight

categories (Figure 7.8) is made.

Figure 7.6 shows that on average children in the Durban-South schools have greater W_{DVar}FEV₁ values than the W_{DVar}FEV₁ values of the children in the schools located in the northern residential areas. Figure 7.7 confirms that indeed the male and female children do not have a common slope nor intercept. Finally, Figure 7.8 shows that the overweight children have the highest W_{DVar}FEV₁ values followed by the W_{DVar}FEV₁ values of the children of normal weight.

Chapter 8

Conclusion

The objective of this thesis was to determine if there is a significant relationship between the ambient air pollutant SO₂, the child characteristics and the respiratory condition of the children. To achieve this, the model-based approach was used. After exploring the mixed model and the hierarchical model, the longitudinal mixed model was used for the analysis of the survey data. To investigate the fixed effects mean structure, individual and mean profile plots of the dependent variable were considered. But since interpretation of these plots was not easy, the loess fit was considered. The aim of these non-parametric curves was to smoothen the profile plots in order to make the choice of the fixed-effects structure easier. To check for multicollinearity in the independent variables, correlations and the variance inflation factor (VIF) were used. These checks showed the presence of multicollinearity. To solve for the problem of multicollinearity, the principal component analysis (PCA) was considered as the remedy.

The existence of multicollinearity between the two summary pollutant measures led to the fitting of two separate models for each of the pollutant measures. The model with the peaks as the independent pollutant measures was referred as the 'peaks pollutant model', and the model with the 8-hour maximum as independent pollutant measures

was referred to as the '8-hour maximum pollutant model'. Since the pollutant summary measures were correlated, the best option would have been to drop one of summary measures and fit only one model with the preferred summary measure. However, it was of interest to see if the two pollutants have the same effect on the children's respiratory condition.

The existence of serial correlation by use of a scatter plot of residuals was explored. A constant variance was also assessed for by use of the variance function. The compound symmetric (CS), autoregressive order one (AR(1)) and the unstructured (UN) covariance structures were considered in the model selection. The compound symmetric structure was selected as being the most appropriate covariance structure for the error term. For the random effects, the unstructured (UN) covariance structure was found to be the most appropriate. For the fixed effects inference, the likelihood ratio (LR) test was used. For the random effects inference, a method used by Stram and Lee (1994; 1995) was adopted.

The profiles of the observed and fitted values indicated that model was a good fit. The normality for the random effects and residuals was assessed graphically. Since the histograms showed a symmetric bell-shaped curve, it was concluded that the normality assumption was not violated. The scatter plots for the random effects and residuals were assessed for outlier detection and for any symmetric pattern. All the plots were in favour of the goodness of the fit. Since the survey data was characterized by incompleteness, the application of a modified logistic regression model MAR test resulted in the conclusion that the survey data had an MAR type of missingness.

In general, school, sex and weight were found to have a significant effect on the respiratory condition of the children. Despite the fact that schools in Durban-South recorded higher pollutant measures as compared to those located in the northern residential areas, the children attending school in the northern residential areas were

found to have poorer respiratory conditions as compared to those attending school in Durban-South.

The male children and female children were found to differ in their respiratory conditions. This is in line with the conclusions drawn by Buist (1982) and Schwartz, Katz, Fegley and Tockman (1988a, 1988b) that that girls are found to have higher expiratory flows than boys after correcting for body size.

The overweight children and children of normal weight; and underweight children and children of normal weight differ in their respiratory conditions. The poorest respiratory conditions were registered by the overweight children. This falls in line with the conclusions drawn by Schoenberg, Beck and Bouhuys (1978) and Dockery, *et al.* (1985) that lung function is decreases at both extremes of weight.

The model also showed no evidence that a rise in the pollutant measures caused a significant deterioration in the respiratory condition of the children in the study. Instead, the results showed that a five day prior exposure to an increase in the pollutant levels led to an improvement in the respiratory conditions of the children. Results from the Settlers primary school health study (Robins *et al.*, 2002), however, showed that prior one day and/or two day exposure to increasing levels of the pollutant significant resulted in the deterioration of the respiratory conditions of the children considered. The difference in the results may be attributed to the fact that not all the survey factors considered in the Settlers primary school health study were considered in this thesis.

The 8-hour maximum pollutant model offers the same conclusions as the peaks pollutant model which means that the two pollutant summary measures do provide similar results.

Further study is needed on the development of MAR tests specifically for longitudi-

nal data with an intermittent mode of missingness. In addition, future research should aim at including more child factors in the model. These factors include blood tests and their exposure to cigarette smoke. Other factors that could be included in the model would be the housing and family conditions.

Bibliography

Abbey, D. E., Petersen, F., Milis, P. K., and Beeson, W. L. (1993). Long-term ambient concentration of total suspended particulates, ozone and sulfur dioxide and respiratory symptoms in nonsmoking population. *Archives of Environmental Health*, 48, 33-46.

Altham, P.M.E. (1984). Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society, Series B*, 46, 118-119.

Baker, S.G. and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to non-ignorable non-response. *Journal of the American Statistical Association*, 83, 62-69.

Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical association*, 44, 3, 388-393

Bowerman, L.B., O'Connell, T.R. and Dickey, A.D. (1986). *Linear statistical models : An applied approach*. Duxbury Press:Boston.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical linear models: Ap-*

plication and data analysis methods. Sage: London.

Buist, A.S. (1982). *Evaluation of lung function: Concepts of normality.* In: Simmons DH, ed. *Current pulmonology. Vol 4.* New York: John Wiley and sons.

Buist, A.S. and Vollmer, W.M. (1988). The use of lung function tests in identifying factors that affect lung growth and aging. *Statistics in Medicine*, 7, 11-18.

Cassel, C., Särndale, C. and Wretman, J.H. (1977). *Foundations of inference in survey sampling.* Wiley : New York.

Cochran, W.G. (1977). *Sampling techniques.* Third edition. New York: Wiley.

Cogill, B. (2003). *Anthropometric Indicators measurement guide.* Food and nutrition technical assistance project, Washington, DC: 2003 Revised Edition. <www.fantaproject.org/downloads/pdfs/anthro_2003.pdf>.

Cohen, R.A. (1999). *An Introduction to PROC LOESS for local regression.* Paper 273, SAS users group international conference (SUGI) proceedings. <www.ats.ucla.edu/stat/sasl/library/loessugi.pdf>.

Conaway, M.R. (1992). The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 87, 817-824.

Dalenius, T. (1988). The problems of non-sampling errors. Sampling, *Handbook of statistics*, volume 6, page 41-44.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the royal statistical society, series B*, 39, 1-38.

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-356.

Diggle, P.J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, 45, 1225-1258.

Diggle, P. J., Heagerty, P.J. and Liang, K.Y. (2002). *Analysis of longitudinal data*. Oxford University Press Inc., New York.

Diggle, P. J. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43(1), 49-93.

Diggle, P. J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of longitudinal data*. Oxford science publications. Oxford: Clarenton press.

Dockery, D.W., Ware, J.H., and Ferris, B.G. Jr. (1985). Distribution of forced expiratory volume in one second and forced vital capacity in healthy white

adult never-smokers in six U.S. cities. *American Review of respiratory disease* **131**, 511-520.

Efron, B. and Morris, C.N. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American statistical association*, **74**, 311-319.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, B* **31**, 195-234.

Fanta, S. (2003). Modelling growth in arm circumference of infants in Jimma town, south-west Ethiopia. *Sinet: Ethiop. J. Sci.*, **26(1)**, 1-10.

Ghosh, M. and Meeden, G. (1997). *Bayesian methods for finite population sampling*. London: Chapman and Hall.

Greenless, W.S., Reece, J.S., and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variation being imputed. *Journal of the American Association*, **77**, 251-261.

Hallahar, C. (2004). *Longitudinal data analysis with discrete and continuous responses using Proc MIXED*. <<http://cpcug.org/user/sigstat/PowerPointSlides/ProcMixedPart5.ppt>> .

Hansen, M.H., Hurwitz, W.N. and Madow, W.G.(1953). *Sample survey methods and theory*. New York:Wiley.

Harville, D.A. (1974). Bayesian inference for variance components using error contrasts. *Biometrika*, 61, 383-385.

Hedayat, A. and Sinha, B.K. (1991). *Design and inference in finite population sampling*. Wiley: New York.

Heitjan, D.F. (1997). Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health*, 87(4), 548-550.

Henderson, C.R., Kempthorne, O., Searle, S.R., and C.M. von Krosig, C.N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15, 192-218.

Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills: Sage publications.

Katsouyanni, K., Toulami, G., Spix, C., Schwartz, J., Balducci, F., Medina, S., Rossi, G., Wojtyniak, B., Sunyer, J., Bacharova, L., Schouten, J. P., Ponka, A., and Anderson, H. R. (1997). Short term effects of ambient sulfur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the PLPHEA project. *British Medical Journal* 314 (7095), 1658 - 1663.

Kenward, M.G. and Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983-997.

Kish, L. (1965). *Survey sampling*. Wiley: New York.

Kott, S.P. (June, 2002). *Randomization-assisted model-based survey sampling*. Paper prepared for the fourth Biennial international conference of statistics, probability and related areas. Dekalb, Illinois.

Laird, N.M., Liang, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, 82, 97-105.

Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974

Lessler, J.T and Kalsbeek, W.D. (1992). *Nonsampling Error in surveys*. Wiley: New York.

Levin, B.J. (July, 1999). *The Use of variance component models in the analysis of complex surveys*. Phd thesis: University of Natal.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Little, R.J.A. (1988). Missing data in large surveys. *Journal of business and economic statistics*, 6, 287-301.

Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated stud-

ies. *Journal of the American Association*, **90**, 1112-1121.

Little, J.R. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American statistical association*, **99**, 546-556.

Little, R.J.A. and Rubin, D.A. (1987). *Statistical analysis with missing data*. New York: John Wiley and Sons.

Littell, R.C., Milliken, G. A., Stroup, W.W., and Wolfinger, R.D.(1996). *SAS system for mixed models*. Sas Institute Inc. Cary: NC, USA.

Longford, N.T. (1993). *Random coefficient models*. Clarendon press, Oxford: New York.

Molenberghs, G., Kenward, M.G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika*, **84**, 33-34.

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2001). *Introduction to linear regression analysis: Third edition*. Wiley: New York.

Morrell, C.H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, **54**, 1560-1568.

Neter, J., Wasserman, W. and Kutner, H.M. (1990). *Applied linear statisti-*

cal models: Regression, analysis of variance and experimental designs - third edition. Irwin : Burr Ridge, Illinois.

Neyman, J. (1934). On the two different aspects of representative methods: The method of stratified sampling and method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Park, T. and Brown, M.B. (1994). Models for categorical data with non-ignorable non-response. *Journal of the American Statistical Association*, 89, 44-52.

Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-554.

Pearson, E.S., D'Agostino, R.B., and Bowman, K.O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, 64, 231-246.

Ridout, M.S. (1991). Reader Reaction: Testing for random dropout in repeated measurement data. *Biometrics*, 47(4): 1617-21.

Robins, T., Batterman, S., Lalloo, U., Iruken, E., Naidoo, R., Kistnasamy, B., Kistnasamy, J., Baijnath, N. and Mentz, G. (2002). *Air contaminant exposures, acute symptoms and disease aggravation among students and teachers at the Settlers' School in South Durban*. Interim Report, University of Natal Faculty of Health Sciences/University of Michigan School of Public Health.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R.M. and Herson, J. (1973). Robust estimation in finite population 1. *Journal of the American Association*, 68, 880-889.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Rubin, D.B. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", by M.H. Hasen, W.G. Madow, and B.J. Tepping. *Journal of the American Statistical Association*, 78, 803-805.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D.B. (2004). *Multiple imputation for nonresponse surveys*. New York: John Wiley and Sons.

SAS Institute Inc. (2000), *SAS/STAT Software: Changes and Enhancements, Release 8.1*, Cary, NC: SAS Institute Inc.

<<http://support.sas.com/rnd/app/da/new/801ce/stat/chap6/index.htm>>.

Schoenberg, J.B., Beck, G.J. and Bouhuys, A. (1978). Growth and decay of pulmonary function in healthy blacks and whites. *Respir Physiol* 33, 367-393.

Schwartz, J., Siaster D., Larson, T. V., Pierson, W. E., and Koenig, J. Q.

(1993a). Particulate air pollution and hospital emergency room visits for asthma in Mexico City. *American Review of Respiratory Disease* 147, 826-831.

Schwartz, J. (1993b). Particulate air pollution and chronic respiratory disease. *Environmental Research* 62, 7-13.

Schwartz, J.D., Katz S.A., Fegley, R.W., and Tockman, M.S. (1988a). Analysis of spirometric data from a national sample of healthy 6 to 24 year olds (NHANES II). *American Review of respiratory disease*. 138, 1405-1414.

Schwartz, J.D., Katz S.A., Fegley, R.W., and Tockman, M.S. (1988b). Sex and race differences in the development of lung function. *American Review of respiratory disease*. 138, 1415-1421.

Scott, A.J. and Smith, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64, 830-840.

Searl, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance components*. New York: Wiley.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimates and likelihood ratio tests under nonstandard conditions. *Journal of American Statistical Association*, 82, 605-610.

Sen, A. and Srivastava, M. (1990). *Regression Analysis: Theory, methods and applications*. New York: Springer-Verlag.

Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591-611.

Simon, G.A. and Simonoff, J.S. (1986). Diagnostic plots for missing data in least square regression. *Journal of the American Statistical Association*, 81, 501-509.

Skinner, C.J., Holt D. and Smith, T.M.F. (1989). *Analysis of complex surveys*. New York: Wiley.

Smith, T.M.F. (June, 1999). *Recent developments in sample survey theory and their impact on official statistics. Lecture to commemorate the 25th anniversary of the IASS: Bulletin of the international statistical institute, 52nd session.* Finland.

Smith, T.M.F. (2001). Centenary-Sample surveys. *Biometrika* 88 (1), 167-194.

Snijders, A.B.T. and Bosker, J.R. (1999). *Multilevel Analysis: An introduction to the basic and advanced multilevel modelling*. Sage: London.

Stasny, E.A. (1986). Estimating gross flows using panel data with nonresponse: an example from Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171-1177.

Stram, D.O. and Lee, J.W. (1995). Correction to: Variance components testing in the longitudinal mixed effects model. *Biometrics*, 51, 1196.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by augmentation. *Journal of the American Statistical Association*, 82, 528-550.

Thompson, S.K. (1992) *Sampling*. Wiley: New York.

Thomsen, I. and Tsefu, D. (1988). One the use of models in sampling from finite population. *Handbook of statistics, Volume 6*, 369-397. ES 2250.

Thurston, G. D., Ito, K. Hayes, C. G., Bates, D. V., and Lippmann, M. (May, 1994). Respiratory hospital admissions and summertime haze air pollution in Toronto, Ontario: Consideration of the role of acid aerosols. *Environmental Research*, 65(2), 271-290.

United States Environmental Protection Agency. (April, 1996). *Air quality criteria for particulate matter*. Washington DC: Office of Research and Development. EDA/600/P-95/001aF-001cF.

Valliant, R., Donfman, A.H., and Royall, R.M. (2000). Finite population sampling and inference: A prediction approach. New York: Wiley.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer: New York.

Walters, S., Griffiths, R. K. and Ayres, J. G. (1994). Temporal association between hospital admissions for asthma in Birmingham and ambient level of sulphur dioxide and smoke. *Thorax*, 49, 133-140.

Wolfinger, R.D., Tobias, R.D., and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on scientific computing*, 15(6), 1294-1310.

Wu, M.C. and Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 45, 939-955.

Wu, M.C. and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, 44, 175-188.

Wuensch, K. L. (2004). *Principal component analysis*.

<<http://core.ecu.edu/psyc/wuenschk/MV/FA/PCA.doc>>.

Zewotir, T. and Galpin, J.S. (2004). The behaviour of normal plots under non-normality for mixed models. *South African statistical journal*, 38 (2), 115-136.

Zheng, H. and Little, R. (2003). *Penalized spline non-parametric mixed models for inference about a finite population mean from a two-stage samples*. Working paper series, paper 6: University of Michigan school of public health, <<http://www.bepress.com/umichbiostat/paper 8>>.

Appendix A

Appendix tables

Table A.1: The peak counts of the SO₂ measurements on a particular date per school

Date	Dirke Uya	Fetret			
5/26/2004	0	0	0	0	0
5/27/2004	0	0	0	0	0
5/28/2004	5	0	0	0	0
5/29/2004	0	1	0	0	0
5/30/2004	0	5	0	0	0
5/31/2004	0	3	0	0	0
6/1/2004	0	0	0	0	0
6/2/2004	0	0	0	0	0
6/3/2004	0	0	0	0	0
6/4/2004	0	0	0	0	0
6/5/2004	0	0	0	0	0
6/6/2004	0	0	0	0	0
6/7/2004	0	0	0	0	0
6/8/2004	0	0	0	0	0
6/9/2004	0	0	0	0	0
6/10/2004	0	0	8	0	8
6/11/2004	0	10	0	11	0
6/12/2004	17	0	0	0	0
6/13/2004	0	0	0	0	0
6/14/2004	2	0	0	0	0
6/15/2004	0	0	0	0	0
6/16/2004	0	0	0	0	0
6/17/2004	0	0	0	0	0
6/18/2004	0	0	0	0	0
Average	1	0.781666667	0.333333333	0.458333333	0.25
Total	24	19	8	11	6

Table A.2: The 8-hour maximum of the SO₂ moving averages on a particular date per school

Date	Primary schools				
	Nizam	Assegal	Dirkie Uys	Femdale	Ngazana
5/26/2004	10.1536	24.4239	5.5709	5.5833	2.3555
5/27/2004	28.0677	62.9596	7.0601	2.5313	3.1208
5/28/2004	85.2658	64.6304	13.9535	4.6413	6.2136
5/29/2004	44.9198	83.2650	63.1702	10.3034	11.9144
5/30/2004	11.1007	90.6135	41.5627	1.4493	1.7273
5/31/2004	9.1771	105.6473	7.4935	2.4462	3.1358
6/1/2004	11.3828	27.6030	11.8914	4.5208	4.0098
6/2/2004	11.7906	26.8658	4.9856	6.0521	3.5540
6/3/2004	43.9071	34.9979	8.8470	3.8854	2.6772
6/4/2004	33.1027	44.7340	15.0948	9.5833	6.8086
6/5/2004	40.2542	54.2685	24.9118	12.6563	6.8044
6/6/2004	33.8925	35.2094	12.6995	10.0521	6.9992
6/7/2004	27.9890	34.7664	19.0134	5.6979	5.0569
6/8/2004	29.9129	37.7935	24.3106	6.9479	7.0019
6/9/2004	10.6843	45.7597	23.7021	5.8316	2.0722
6/10/2004	8.4855	22.0854	90.1278	5.2211	78.5806
6/11/2004	17.2866	128.4613	20.7058	135.8172	5.0334
6/12/2004	211.2263	38.1902	21.2300	13.6563	7.5739
6/13/2004	40.4152	39.0975	18.5814	4.3125	4.4525
6/14/2004	59.5694	33.8535	23.8160	6.3333	6.8632
6/15/2004	41.4365	34.0411	22.3100	6.6563	6.5275
6/16/2004	14.7027	20.3842	12.3301	13.0000	3.7753
6/17/2004	10.4485	30.2914	11.0384	13.0000	5.7555
6/18/2004	23.6675	23.4636	11.8282	4.4688	2.5859
Average	35.7850	47.6419	21.5098	12.2770	8.1083
Total	858.8390	1143.4063	516.2348	294.6474	194.5992

Table A.3: The SO₂ levels above the standard guideline of 191pb as well as the date, time and the school in which the high SO₂ level occurs

School	Date	Time	SO ₂	School	Date	Time	SO ₂	School	Date	Time	SO ₂
Nizam	5/27/2004	2200	197	Assegai	5/28/2004	810	199.9	Dirkie Uys	6/9/2004	1440	216.7
	5/27/2004	2310	213.2		5/29/2004	1150	211.7		6/9/2004	1450	677
	5/27/2004	2320	469.7		5/29/2004	1200	208.1		6/9/2004	1500	684.3
	5/27/2004	2330	364.2		5/29/2004	1230	202.5		6/9/2004	1510	684.4
	5/27/2004	2340	289		5/29/2004	1250	200.7		6/9/2004	1520	573.5
	6/11/2004	1120	315.1		5/29/2004	1320	206.3		6/9/2004	1530	465.3
	6/11/2004	1130	569.7		5/30/2004	1120	232		6/9/2004	1540	434.8
	6/11/2004	1140	591.9		5/30/2004	1140	202.5		6/9/2004	1550	287.8
	6/11/2004	1150	356.3		5/30/2004	1210	220.4	Ferndale	6/10/2004	1015	638.5
	6/11/2004	1200	577.6		6/10/2004	1450	483		6/10/2004	1025	652
	6/11/2004	1210	611.5		6/10/2004	1500	700		6/10/2004	1035	664.5
	6/11/2004	1220	621		6/10/2004	1510	656.9		6/10/2004	1045	673
	6/11/2004	1230	628.5		6/10/2004	1520	548.5		6/10/2004	1055	678.5
	6/11/2004	1240	636.2		6/10/2004	1530	526.9		6/10/2004	1105	596
	6/11/2004	1250	659.6		6/10/2004	1540	464.8		6/10/2004	1115	542.5
	6/11/2004	1300	687.8		6/10/2004	1550	399.9		6/10/2004	1125	502.5
	6/11/2004	1310	685.1		6/10/2004	1600	427.1		6/10/2004	1135	452
	6/11/2004	1320	586.3		6/10/2004	1610	712		6/10/2004	1145	352
	6/11/2004	1330	542.9		6/10/2004	1620	465.9	Ngazana	6/9/2004	1610	471.3
	6/11/2004	1340	483.7						6/9/2004	1620	730
	6/11/2004	1350	400.5						6/9/2004	1630	757
	6/11/2004	1400	345.8						6/9/2004	1640	715
	6/13/2004	1840	351.5						6/9/2004	1650	392.2
	6/13/2004	1850	436.9						6/9/2004	1700	475.4