Longitudinal Analysis of the effect of meteorological factors, allergens, and air pollution on respiratory condition in children

Longitudinal Analysis of the effect of meteorological factors, allergens, and air pollution on respiratory condition in children

By

Yunna Song, B.Sc.

A Project
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2007)
(Statistics)

McMaster University
Hamilton, Ontario

TITLE:	Longitudinal Analysis of the effect of meteorological factors, allergens, and air pollution on respiratory condition in children
AUTHOR:	Yunna Song, B.Sc (McMaster University, Canada)
SUPERVISOR:	Professor Aaron Childs
NUMBER OF PAGES:	x, 54

Abstract

In this report we explore how the effect of meteorological factors, allergens, and air pollution on respiratory conditions in children using longitudinal data. Our analysis makes use of a dataset from the DAVIS study in southern Ontario. The response variables are children's lower respiratory tract (URT) and upper respiratory tract (URT) scores. The explanatory variables are readings of various meteorological, allergen, and air pollution factors. First we make use of generalized estimating equations to find the main factors that are associated with certain respiratory conditions in children as measured by LRT and URT scores. Then we determine whether there are any interactions between the significant factors associated with LRT/URT scores. Comparisons between case and control groups are made to determine whether children with asthma are more sensitive to any of the changes in meteorological, allergen, and air pollution factors. The analysis results show that the significant factor that is associated with LRT scores for children with asthma is the two-day lag daily average changes in air pressure. On average an increase in air pressure will result in an increase in children's LRT scores. The interaction terms that remained in the final model show some degree of significance but without strong evidence to support it. Children in the case groups are more sensitive to meteorological factors, allergens, and air pollution than the children in control groups.

Acknowledgements

I would like to thank my supervisor, Dr. Aaron Childs, for his great guidance, support and patience, and for giving me an opportunity to work on this project. I really appreciate Dr. Childs' supervision through my project.

I would like to thank Mr. Neil Johnston and Jennifer Dai for supplying the data used in this project.

I would like to thank Dr. Peter Macdonald and Dr. Rong Zhu for being the examiners of this project. I appreciate all the comments and criticism. I also would like to thank them for their help and support during my two-year study.

I would like to thank my parents, my husband, my brother and sister-in-law for their constant support and encouragement throughout my academic career.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Asthma is the leading cause of hospitalization in children. In Canada, 10 to 15 percent of children are reported to have asthma, though it is believed that the rate could be as high as 20 percent. It is believed that meteorological factors, air pollution, and allergens may exacerbate asthma attacks in children. But few studies have used longitudinal analysis to investigate the relationship among lower/upper respiratory tract (LRT/URT) problems, meteorological factors, air pollution, and allergens on a daily basis, especially the effect of any combination of meteorological factors, air pollution and allergens. Many studies have examined the daily count of asthma admission or emergency room visits in relation to short term fluctuations in measured thunderstorms, pollen, trees, and air pollution. Furthermore, the etiology in asthmatic children is unknown. Knowing the association among meteorological factors, allergen and air pollution on asthma in children can help the prevention of asthma attacks.

1.2 DAVIS Study Asthma Data

The data for the current study were obtained from Mr. Neil Johnston, a research member from the Firestone Institute for Respiratory Health at St.Joseph's Hospital in Hamilton, Ontario. The individual personal file came in as an Access file and the meteorological, air pollution, and allergens data came in as Excel files. About 5% of the subjects have missing values on the skin sensitivity test.

1.2.1 Lower and Upper Respiratory Tract Score Data

The individual data: Daily scores of LRT and URT were collected from July 2003 to December 2004 inclusive in Southern Ontario using a Secure Internet based system from a cohort of 208 asthmatic and non-asthmatic children. The lower respiratory tract consists of the part of the respiratory system including the trachea, primary bronchi, and lungs. The upper respiratory tract consists of the part of the respiratory system including the nasal cavity, pharynx, and larynx. The scores are given according to the severity of the symptoms in each part. The LRT scores are in the range of 0 to 18 and the URT scores are from 0 to 9. Lower respiratory tract problems are generally more serious than upper respiratory tract. Children's demographic characteristics such as age and gender were collected. Children were categorized into different groups according to treatment/control, atopic/non-atopic, asthmatic/non-asthmatic, gender, and age. The asthmatic group consists of those subjects that have had physician confirmed asthma attacks in the past. The atopic group consists of subjects that failed at least two types of skin sensitivity test. Subjects in the study were between ages of 5 and 11.

1.2.2 Air pollution, Aeroallergen, and Meteorological Data

The weather data: Hourly measures of temperature, pressure and humidity were recorded every day from November 2003 to December 2004. Then daily average temperature, humidity and pressure were computed as the mean of the three respective values. The difference between daily maximum and minimum were calculated as the range for temperature, pressure and humidity. The temperature, pressure and humidity change were computed as the difference of average temperature, pressure and humidity and the one day lag values of the variables respectively. The absolute daily changes of the values were computed as well. Also whether a thunderstorm has occurred hourly was recorded. Here, we are interested in the number of hours in a day which a thunderstorm has occurred. Then each of the calculated variables were plotted against time to check whether there is any trend. Moving average were used to remove any trend or seasonality. From the plots, only the mean temperature need a moving average. Then the one day lag and two day lag values were also computed to observe the day to day changes. Finally the logarithms values were computed for each of the variables.

The air pollution data: Hourly measures of Sulphur Dioxide (SO_2), Nitric Oxide (NO), Nitrogen Dioxide (NO_2), Nitrogen Oxides (NO_x), Carbon Monoxide (CO), and Ozone (O_3) were collected from January 2004 to December 2004. There were five monitoring sites for pollutants. We used Brantford, Burlington, and Hamilton downtown since these three sites have similar amounts of the pollutants we were interested in. The corresponding average and range were computed for analysis. If any reading from one site is greater than two times of the average value of the other two sites, then the average will be used in the final analysis. No moving averages were needed for the pollutant data. Finally the corresponding logarithms for the pollutant variable and their one day lag and two day lag values were computed.

The allergens data: There are no allergen readings in the winter time. The aeroallergen data has been collected from April 2004 to November 2004 using rotational impaction sampling located in each city. The particle adhering to the silicon grease-coated sample rods was analyzed to determine the number of particles present per cubic meter of air sampled. Daily measures of total trees, total weeds, total pollen, total basidiomycetes, total ascomycetes, and total other spores were collected from Hamilton, London and Brampton sites. Since there were quite a lot of missing values in the Hamilton site, the other two site readings were used to fill the values. First the average of the other two sites with respect to each of the six variables are calculated. If the reading in Hamilton site is missing, then the average of London and Brampton value will be used; if the value in Hamilton value is not missing, but the reading is two times greater than the average of the other two sites, then the average of London and Brampton readings will be used. Since all the aeroallergen variables have trend, moving averages are done for each of them. Then the corresponding logarithms for each variable and their one day lag and two day lag values were computed.

1.2.3 Missing Data

It is desirable to have complete data for the longitudinal analysis. The method we used to estimate the missing values was to use the available measurements from the other monitoring sites on the same day. Most of the missing values occurred in the air pollutant data. For example, if the pollutant measure was missing from the Hamilton downtown site, then the average of readings from the other two sites was used. We were also informed that the pollutant measures from the Hamilton downtown site were not reliable. So we also checked to see if the Hamilton pollutant readings were within a reasonable range. If a value is missing completely for all the sites on a particular day (which happened with some of the aeroallergen data), then the interpolation method was used to fill in the missing values. The method used was to fit a cubic.

1.3 Objectives

The purpose of this study is to test whether changes in children's LRT or URT scores are associated with meteorological factors, allergens, and air pollution among children with different characteristics. Also we want to test whether the interaction between these factors are significant for predicting LRT or URT scores.

1.4 Descriptive Analysis of DAVIS Data

Table 1.1 below gives a listing and description of all the variables in the study.

Table 1.1: *Description of the original main variables*

Variable name	**Description**
Char1	Age group
Char2	Gender
Group1	Asthmatic and Non-Asthmatic
Group2	Atopic and Non-Atopic
Group3	adjudicated Asthmatic and adjudicated Non-Asthmatic
LRT	Lower respiratory tract score
URT	Upper respiratory tract score
Temperature	Hourly temperature of the designated city
Pressure	Hourly pressure of the designated city(kPA)
Humidity	Hourly humidity of the designated city
Thunderstorm	Has a thunderstorm occurred within the last hour
Total pollens	Daily counts of total pollens (spores/m3)
Total weeds	Daily counts of total weeds (spores/m3)
Total trees	Daily counts of total trees (spores/m3)
Total ascomycetes	Daily Counts of total ascomycetes (spores/m3)
Total basidiomycetes	Daily counts of total basidiomycetes (spores/m3)
Total spores and other	Daily counts of total spores and other (spores/m3)
Sulphur Dioxide	Hourly measures of Sulphur Dioxide (ppb*)
Nitric Oxide	Hourly measures of Nitric Oxide (ppb*)
Nitrogen Dioxide	Hourly measures of Nitrogen Dioxide (ppb*)
Nitrogen Oxides	Hourly measures of Nitrogen Oxides (ppb*)
Carbon Monoxide	Hourly measures of Carbon Monoxide (ppm**)
Ozone	Hourly measures of Ozone (ppb*)

* parts per billion

** parts per million

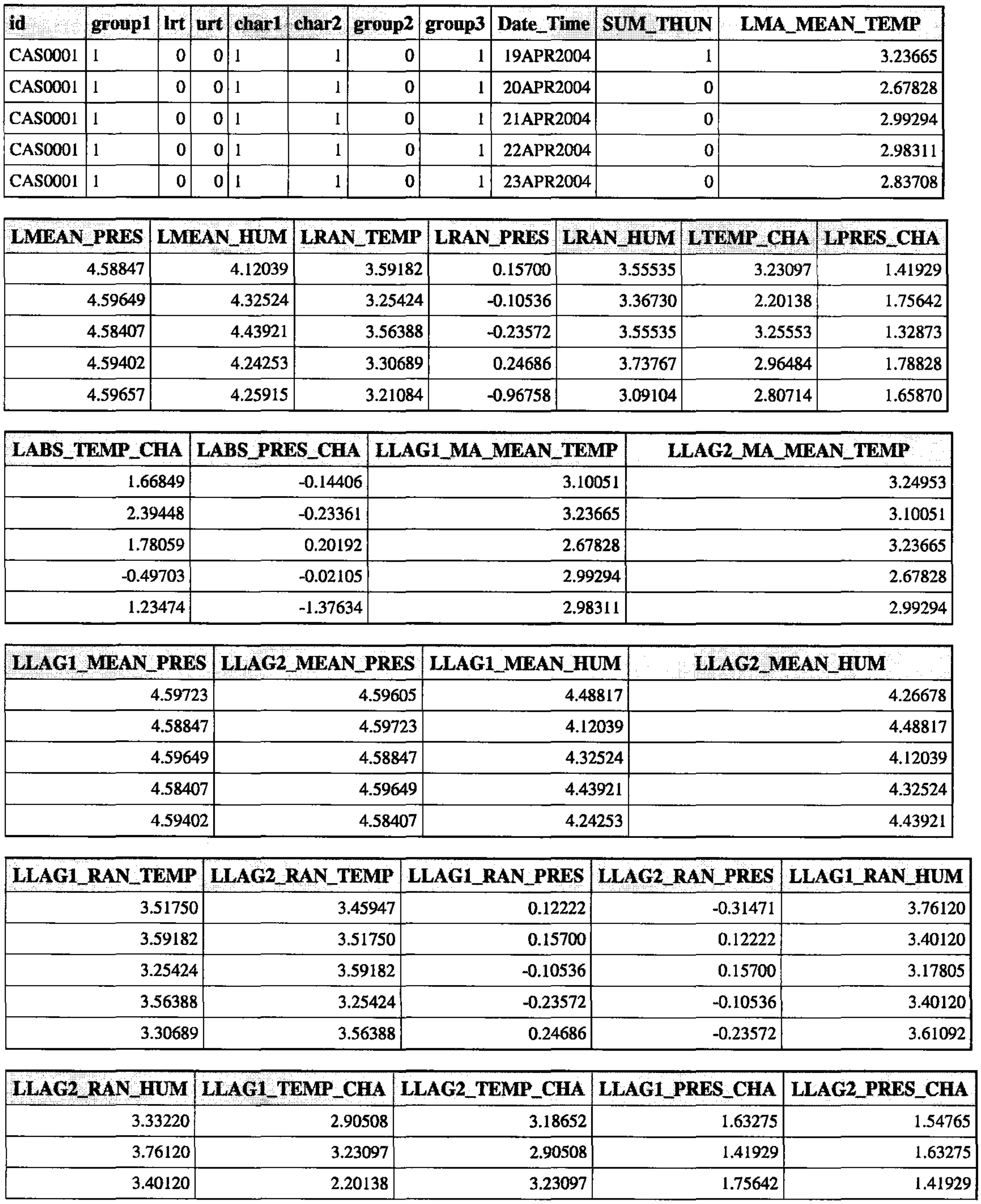
There were a total 208 children in the study. From table 1.2 we see that 60 percent of the subjects were male and 40 percent were female. A little more than 70 percent of the children were 5-8 years of age. In group 2, 17 of the subjects didn't have a skin test. The difference between group 1 and group 3 is that group 3 is with physician confirmed asthmatic and non-asthmatic cases, where group 1 is without confirmed examination.

Table 1.2: *Frequency of subjects' characteristics*

Group	Variable Name	Freq	Percent
Char1	Male	121	58.17
	Female	87	41.83
Char2	Age 5-8	149	71.63
	Age 9-11	59	28.37
Group1	Asthmatic	125	60.10
	Non-Asthmatic	83	39.90
Group2	Atopic	139	72.77
	Non-Atopic	52	27.23
Group3	Adjudicated Asthmatic	147	70.67
	Adjudicated Non-Asthmatic	61	29.33

Each subject was assigned a unique identifier. In cross sectional data, each subject has one row. In longitudinal data, each subject may have several rows, where each row represents one measurement. Unlike repeated measures, the measurements are correlated. Here, each subject has about 200 rows. Each row corresponds to daily measurement of the same subject. Figure 1.1 shows an example of data for one subject for the dates April 19-23, 2004, as well as the logarithm of the meteorological data for these days. Figure 1.2 shows typical Plots of individual Lower respiratory tract (LRT) scores from day to day for a typical asthmatic and non-asthmatic individual. Note that there are lots of LRT score variations in the asthmatic group Vs. a few variation in Non-Asthmatic group.

Figure 1.1: *Snapshot of one individual record*



id	group1	lrt	urt	char1	char2	group2	group3	Date_Time	SUM_THUN	LMA_MEAN_TEMP
CAS0001	1	0	0	1	1	0	1	19APR2004	1	3.23665
CAS0001	1	0	0	1	1	0	1	20APR2004	0	2.67828
CAS0001	1	0	0	1	1	0	1	21APR2004	0	2.99294
CAS0001	1	0	0	1	1	0	1	22APR2004	0	2.98311
CAS0001	1	0	0	1	1	0	1	23APR2004	0	2.83708

LMEAN_PRES	LMEAN_HUM	LRAN_TEMP	LRAN_PRES	LRAN_HUM	LTEMP_CHA	LPRES_CHA
4.58847	4.12039	3.59182	0.15700	3.55535	3.23097	1.41929
4.59649	4.32524	3.25424	-0.10536	3.36730	2.20138	1.75642
4.58407	4.43921	3.56388	-0.23572	3.55535	3.25553	1.32873
4.59402	4.24253	3.30689	0.24686	3.73767	2.96484	1.78828
4.59657	4.25915	3.21084	-0.96758	3.09104	2.80714	1.65870

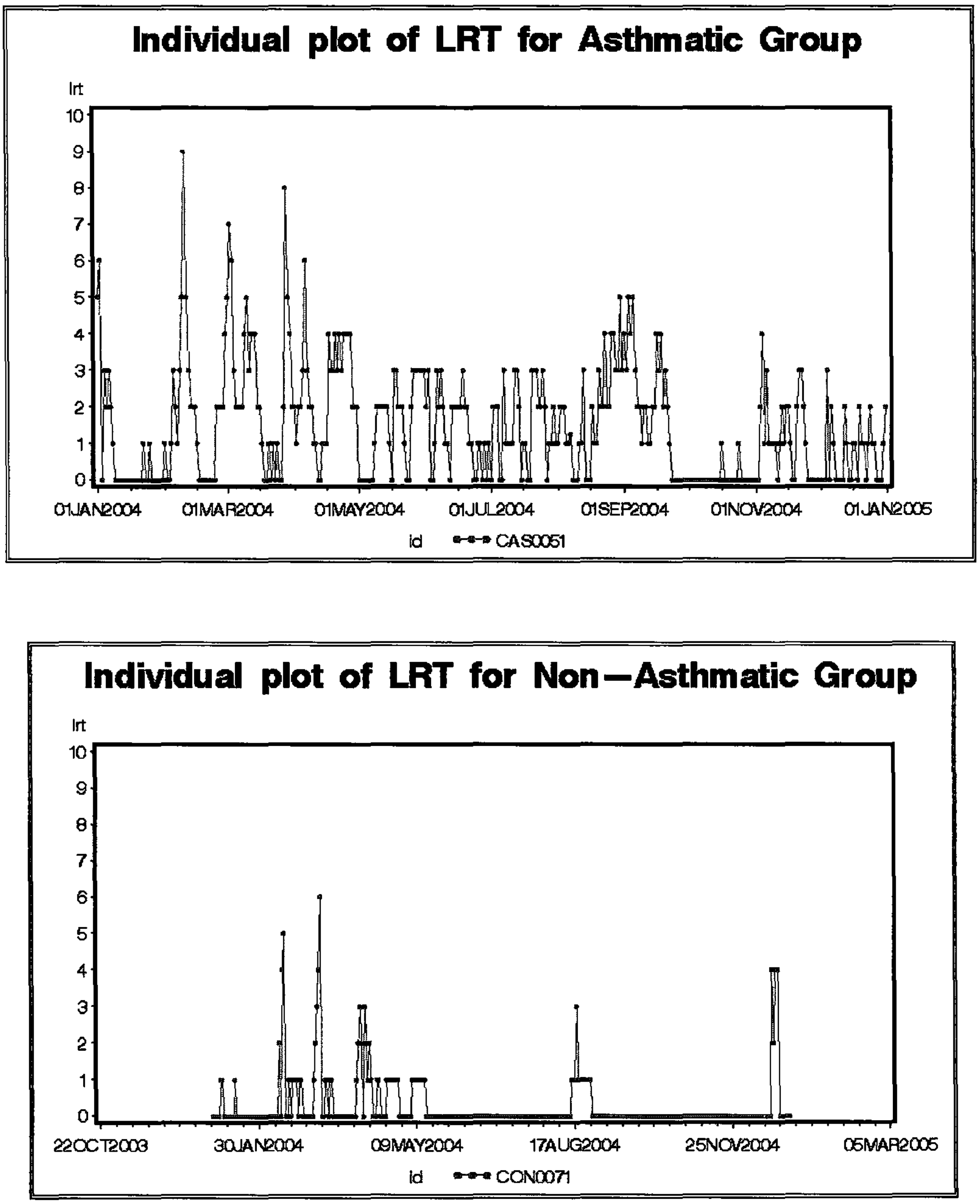
LABS_TEMP_CHA	LABS_PRES_CHA	LLAG1_MA_MEAN_TEMP	LLAG2_MA_MEAN_TEMP
1.66849	-0.14406	3.10051	3.24953
2.39448	-0.23361	3.23665	3.10051
1.78059	0.20192	2.67828	3.23665
-0.49703	-0.02105	2.99294	2.67828
1.23474	-1.37634	2.98311	2.99294

LLAG1_MEAN_PRES	LLAG2_MEAN_PRES	LLAG1_MEAN_HUM	LLAG2_MEAN_HUM
4.59723	4.59605	4.48817	4.26678
4.58847	4.59723	4.12039	4.48817
4.59649	4.58847	4.32524	4.12039
4.58407	4.59649	4.43921	4.32524
4.59402	4.58407	4.24253	4.43921

LLAG1_RAN_TEMP	LLAG2_RAN_TEMP	LLAG1_RAN_PRES	LLAG2_RAN_PRES	LLAG1_RAN_HUM
3.51750	3.45947	0.12222	-0.31471	3.76120
3.59182	3.51750	0.15700	0.12222	3.40120
3.25424	3.59182	-0.10536	0.15700	3.17805
3.56388	3.25424	-0.23572	-0.10536	3.40120
3.30689	3.56388	0.24686	-0.23572	3.61092

LLAG2_RAN_HUM	LLAG1_TEMP_CHA	LLAG2_TEMP_CHA	LLAG1_PRES_CHA	LLAG2_PRES_CHA
3.33220	2.90508	3.18652	1.63275	1.54765
3.76120	3.23097	2.90508	1.41929	1.63275
3.40120	2.20138	3.23097	1.75642	1.41929

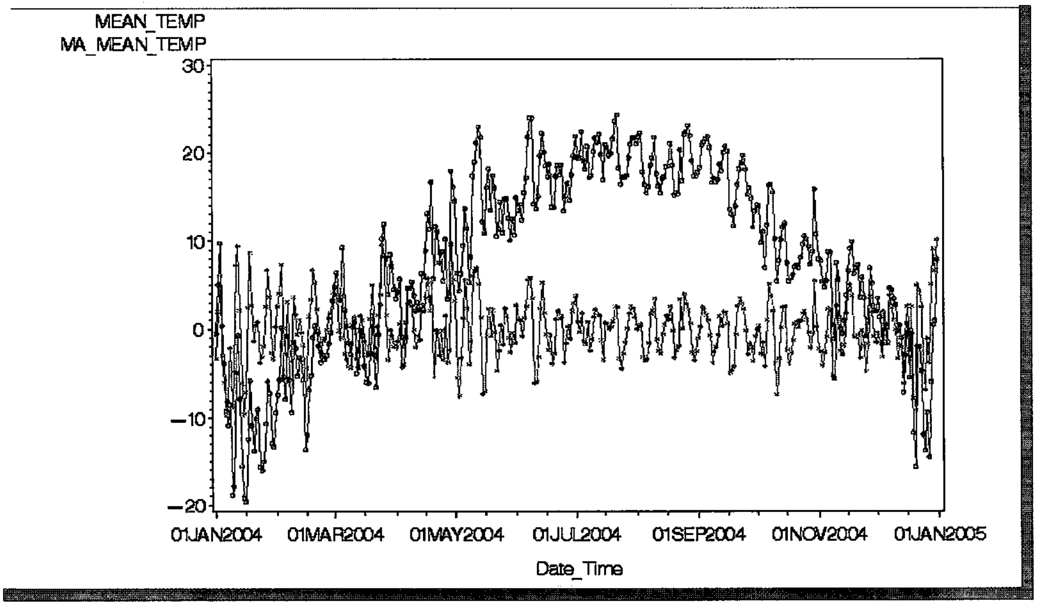
Figure 1.2: *LRT Scores for Asthmatic and Non-Asthmatic Groups*



1.5 Preliminary Analysis

The weather data are stored monthly. A macro has been programmed to import all the weather data at once (Appendix A). We first plotted the independent variables to check for the trend and seasonality. When a trend was present we applied a difference of daily average (Appendix B) to the variable to eliminate the trend. For example, Figure 1.3 shows that a lag average is needed for average temperature (MEAN-TEMP), but not for range temperature (RAN-TEMP). Similarly, Figure 1.4 shows that a lag average is not needed for daily average of nitrogen dioxide and sulphur dioxide.

Figure 1.3: *Plots of daily average temperature, range temperature and their lag average Vs. Date time.*



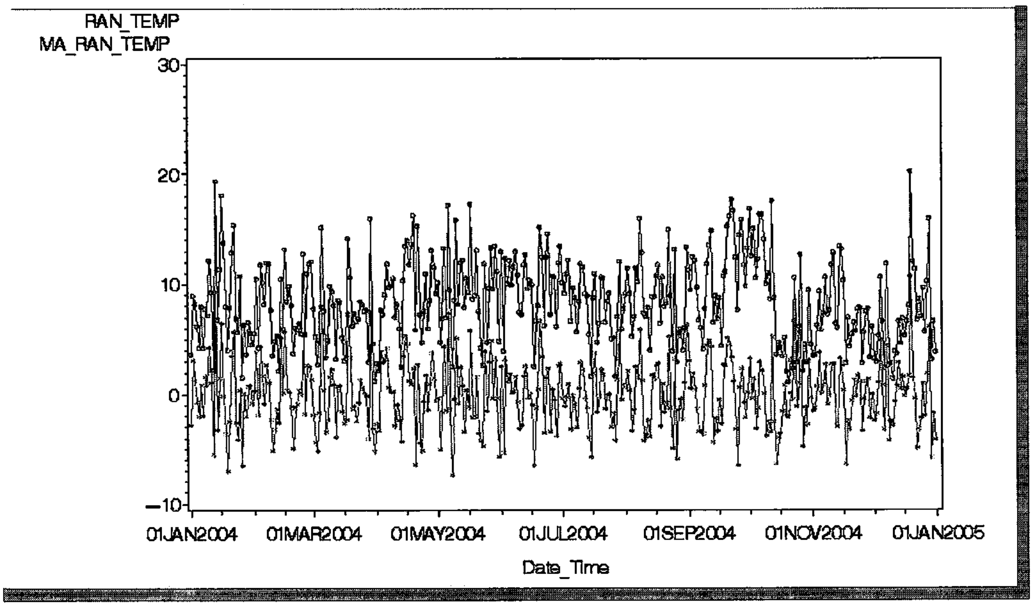
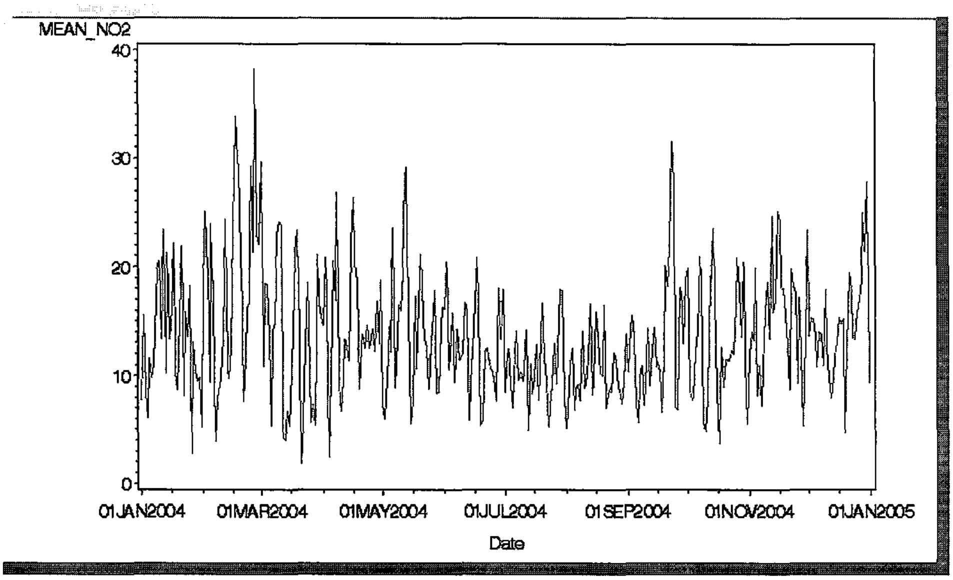
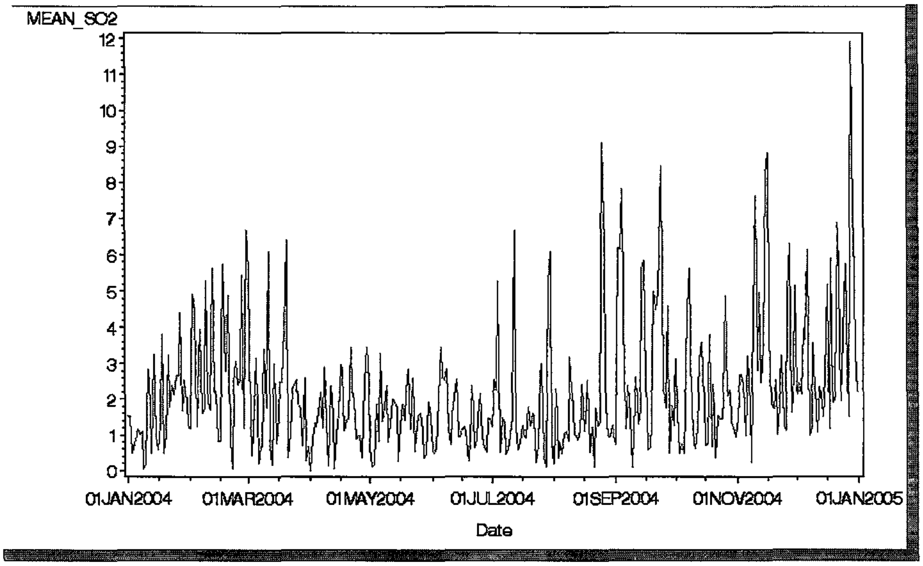


Figure 1.4: *Plots of daily average of Nitrogen Dioxide and Sulphur Dioxide Vs. Date*





Chapter 2

Generalized Linear Models

A generalized linear model is a generalization of linear models. A linear model specifies a linear relationship between the dependent variable Y and a set of of independent predictors $X's$. However, many relationships can't be summarized in a simple linear equation and normality can't be assumed. The first reason is that the effect of predictors on dependent variable may not be linear. The second reason is that the dependent variable of interest has a non-continuous distribution.

2.1 General Assumptions

Generalized Linear Models are an extension of the linear modeling process that allows models to be fit to data that follow probability distributions other than the normal distribution, such as the poisson, binomial, multinomial, etc. In the analysis of generalized linear model, we need to specify three parts: the distributional assumption, the systematic component, and the link function.

2.1.1 Generalized linear models

The classical linear regression models for scalar response Y_j and k covariates $x_{j1}, ..., x_{jk}$ is usually written as

$$Y_j = \beta_0 + \beta_1 x_{j1} + ... + \beta_k x_{jk} + \epsilon_j \tag{2.1}$$

or, defining $\boldsymbol{x_j}= (1, x_{j1}, ..., x_{jk})'$, $\quad Y_j = \boldsymbol{x_j}'\boldsymbol{\beta} + \epsilon_j \quad \boldsymbol{\beta}= (\beta_0, ..., \beta_k)'$,

The Y_j's are assumed to be independent. When the response is continuous, it is often assumed that the ϵ_j's are independent $N(0, \sigma^2)$, so that

$$Y_j \sim N(\boldsymbol{x_j}'\boldsymbol{\beta}, \sigma^2). \tag{2.2}$$

That is, the classical, normal-based regression model may be summarized as:

- **Mean**: $E(Y_j) = f(\boldsymbol{x_j}'\boldsymbol{\beta})$.
- **Probability distribution**: Y_j follow a normal distribution for all j and are independent.
- **Variance**: $\text{var}(Y_j) = \sigma^2$(constant variance).

2.1.2 Generalization

For response variables that are not well represented by a normal distribution, the above model is no longer appropriate. A generalized linear model extends the classical linear regression model as follows.

- The mean of Y_j is assumed to be in the form

  $$E(Y_j) = f(\boldsymbol{x_j}'\boldsymbol{\beta}), \tag{2.3}$$

  where the function f is monotone and differentiable. This means that there is a unique function g, called inverse function of f, such that we may re-express the model in the form

  $$g(E(Y_j)) = \boldsymbol{x_j}'\boldsymbol{\beta}. \tag{2.4}$$

The function g is called the link function, because it "links" the mean and covariates. The linear combination of covariates and regression parameters $\boldsymbol{x_j}'\boldsymbol{\beta}$ is called the linear predictor.

- The probability distribution Y_j is assumed to be one of the scaled exponential family class.

- The variance of Y_j is assumed to be the form dictated by the distribution:

$$\mathrm{var}(Y_j) = \phi V(E(Y_j)), \tag{2.5}$$

where the function $V()$ depends on the distribution and ϕ might be equal to a known constant. The function V is referred to as the variance function. The parameter ϕ is often called the dispersion parameter.

2.2 Iterative Reweighted Least Squares

A natural approach to estimate $\boldsymbol{\beta}$ in all generalized linear model is to use maximum likelihood. The ML estimator for $\boldsymbol{\beta}$ is a solution to

$$\sum_{j=1}^{n} \frac{1}{V(f(\boldsymbol{x_j}'\boldsymbol{\beta}))} \{Y_j - f(\boldsymbol{x_j}'\boldsymbol{\beta})\} f'(\boldsymbol{x_j}'\boldsymbol{\beta})\boldsymbol{x_j} = 0, \tag{2.6}$$

where $f(\boldsymbol{x_j}'\boldsymbol{\beta})$ is the mean of Y_j and f is the inverse link function. Here, $\boldsymbol{\beta}$ appears as a function of f. Unlike with ordinary least squares, it is not possible to solve the above equation for $\boldsymbol{\beta}$ explicitly,

$$\sum_{j=1}^{n} (Y_j - \boldsymbol{x_j}'\boldsymbol{\beta})\boldsymbol{x_j} = 0. \tag{2.7}$$

Therefore we must use a numerical algorithm.

The algorithm used to solve this equation is called Iteratively Reweighted Least Squares, and is performed as follows.

- Give a starting value $\boldsymbol{\beta}^{(0)}$ for $\boldsymbol{\beta}$, : Evaluate the weight at $\boldsymbol{\beta}^{(0)}$: $1/V\{f(\boldsymbol{x_j}, \boldsymbol{\beta}^{(0)}\}$

- Pretending the weights are fixed constants not depending on $\boldsymbol{\beta}$, solve equation (2.6). This still requires a numerical technique, but maybe accomplished by something that is approximately like solving (2.7). This gives a new value $\boldsymbol{\beta}^{(1)}$.
- Evaluate the weights at $\boldsymbol{\beta}^{(1)}$, repeat until two successive $\boldsymbol{\beta}^{(1)}$ are very close.

2.3 Sampling Distribution of the MLE

The sampling distribution of the estimator $\hat{\boldsymbol{\beta}}$ can not be derived in closed form. By large sample theory, when n is large, the IRWS/ML estimator satisfies,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \phi(\Lambda' V^{-1} \Lambda)^{-1}). \tag{2.8}$$

- Λ is a $(n \times p)$ matrix whose (j, k) element $(j = 1, ..., n, k = 1, ..., p)$ is the derivative of $f(\boldsymbol{x_j}'\boldsymbol{\beta})$ with respect to the kth element of $\boldsymbol{\beta}$.
- V is $(n \times n)$ diagonal matrix with diagonal element $V\{f(\boldsymbol{x_j}'\boldsymbol{\beta})\}$.

2.4 Hypothesis Testing

It is common to use the Wald testing procedures to test the hypothesis about $\boldsymbol{\beta}$, for the null hypothesis of the form

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}, \tag{2.9}$$

we may approximate the sampling distribution of the estimate $\mathbf{L}\hat{\boldsymbol{\beta}}$ by

$$\mathbf{L}\hat{\boldsymbol{\beta}} \sim N(\mathbf{L}\boldsymbol{\beta}, \mathbf{L}\hat{V}_{\beta}\mathbf{L}'). \tag{2.10}$$

Then if $\mathbf{L}$ is a row vector, we can construct the test statistics and confidence interval from the "z-statistic"

$$z = \frac{\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h}}{SE(\mathbf{L}\widehat{\boldsymbol{\beta}})} \tag{2.11}$$

and more generally, the Wald test statistic would be

$$(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}\widehat{V_{\beta}}L')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h}), \tag{2.12}$$

and compared to the appropriate chi squared critical value with degree of freedom equal to the number of rows.

Chapter 3

Generalized Estimating Equations

Correlated data arise from many health science trials such as case/ control studies with drugs, clinical trials with baseline, and follow up visits and longitudinal studies. The big concern about this kind of data is how to account for the correlated measurements. For longitudinal data for a group of subjects are likely to exhibit correlation between successive measurement, therefore, within subjects, factors are likely to be correlated, but between subjects are likely to be independent. In analysis of correlated data, if the correlation is not taken into account, parameter estimates and standard errors can not be trusted.

The basic ideas of generalized estimating equations (GEE) are introduced by Liang and Zeger (1986). GEE is an extension of generalized linear model that provides a semi-parametric approach to longitudinal data analysis. The GEE methodology models a known function of marginal expectation of dependent variables as a linear function of one or more explanatory variables. With quasi-likelihood methods, statistical models are created by making assumptions about the link function and the relationship about the mean and variance, but without fully specifying the distribution of the response. GEE describes the random component of the model for each marginal response with a common link and variance function.

The GEE methodology provides consistent estimate of the regression coefficients and variance under weak assumption about the actual correlation within a subject. This method relies on the independence among subjects to estimate consistently the variance of the proposed estimators when the assumed working correlation was incorrectly specified. In this section, the basic ideas of GEE are introduced by Liang and Zeger (2002).

3.1 Population-Average Model

The population-average approach is focused on modeling the mean response across the population of units at each time point as a function of time. The model describes how the average across the population of responses at different time points are related over time. Liang and Zeger's (1986) original approach is to forget about trying to model the whole multivariate probability distribution data vector. Instead, the idea is just to model the mean response and the covariance matrix of a data vector as in the normal case. The alternative approach to model fitting for such mean-covariance models for non-normal longitudinal data does not require specification of a full probability model but rather just the mean and covariance matrix.

3.1.1 Mean Variance Model

The mean response model is

$$\mu_{ij} = E(Y_{ij}) = f(x'_{ij}\beta). \tag{3.1}$$

The variance of Y_{ij} is modeled as some function of the mean response μ_{ij}

$$\text{var}(Y_{ij}) = \phi V(\mu_{ij}). \tag{3.2}$$

3.1.2 Overdispersion

Sometime, the model for variance turn out to be inadequate for representing all the variation in observations taken at a particular time across units.

- The aggregate effects of (i) error introduced by taking measurements and (ii) variation because units differ add up to be more than would be expected if we only considered observation on a particular unit.

- There may be other factors involved in data collection that make things look more variable than the usual assumptions might indicate: e.g. the subjects in the Davis study may have not kept accurate records of the number of LRT problems that they experienced during a particular time period, and perhaps recalled it as being greater or less than it actually was.

Therefore, for count data, it is standard to modify the variance model to allow for an additional scale or overdispersion parameter

$$\text{var}(Y_{ij}) = \phi E(Y_{ij}). \tag{3.3}$$

3.1.3 Working Correlation

The last requirement is to specify a model describing correlation among pairs of observations on the same data vector. The model for correlation is attempting to represent how all sources of variation that could lead to association among observations "added up" the aggregate of

- Correlation due to within-subject "fluctuations" on a particular unit (and possibly measurement error).

- Correlation due to the simple fact the observations on the same unit are "more alike" than those from different units.

Some commonly used correlation structures are as follows.

(1) Unstructured correlation: For observations taken at the same time points for different units, this assumption places no restriction on the nature of associations among elements of a data vector. Let Y_{ij} and Y_{ik}, $j, k = 1, ..., n$ be two observations on the same unit where all units are observed at the same time point. If ρ_{jk} represents the correlation between Y_{ij} and Y_{ik}, then $\rho_{jk} = 1$ if $j = k$ and $-1 \leq \rho_{jk} \leq 1$ if $j \neq k$. The implied correlation matrix for a data vector with all n observations is the $n \times n$ matrix

$$\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{n1} & \cdots & \rho_{n,n-1} & 1 \end{pmatrix} \tag{3.4}$$

where $\rho_{jk} = \rho_{kj}$ for all j, k. Thus the unstructured "working" correlation assumption depends on $n(n-1)/2$ distinct correlation parameters.

(2) Compound symmetry (exchangeable) correlation: This assumption says that the correlation between distinct observations on the same unit is the same regardless of when in time the observations were taken. In principle, this model could be used with balanced data, ideally balanced data with missing values, and unbalanced data where time points are different for different units. This structure may be written in terms of a single correlation parameter $0 < \rho < 1$; i.e.

$$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \cdots & \rho & 1 \end{pmatrix} \tag{3.5}$$

(3) One-dependent: This assumption says that only observations adjacent in time are correlated by the same amount $-1 < \rho < 1$. In principle, this model could be used in any situation; however, for unbalanced data with different time points, it may not make sense. The model may be written as

$$\begin{pmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \rho & 1 \end{pmatrix} \tag{3.6}$$

(4) AR(1): This assumption says that correlation among observations "tail of"; if $-1 < \rho < 1$, the model is

$$\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \cdots & \rho^2 & \rho & 1 \end{pmatrix} \tag{3.7}$$

In principle, this model could be used with any situation; however, for unbalanced data with different time points, it may not make sense.

3.1.4 Summary

- Mean-response of a data vector Y_i as a function of time, other covariates and parameter β by using a generalized linear model-type mean structure to represent mean response of each element Y_i.

- Variance of each element Y_i is modeled by the function of the mean that is appropriate for the type of data e.g., count data are taken to have poisson variance structure, which says that the variance of any element of Y_i is the corresponding mean. These models are often modified to allow for greater variation both within and among units by the addition of a dispersion parameter ϕ.

- Correlation among observations on the same unit is represented by choosing a model, such as the correlation structures corresponding to AR(1), one-dependent or other specification. Because there is some understanding in doing this and no formal way to check it, the chosen model is referred to as the "working" correlation matrix.

With the above considerations, the mean response model and variance of Y_{ij} is the same as (3.1) and (3.2). The standard deviation of Y_{ij} is given by $\{\phi V(\mu_{ij})\}^{1/2}$. Suppose unit i has n_i observations, so that $j = 1, ..., n_i$. Define the standard deviation matrix for unit i as the $(n_i \times n_i)$ diagonal matrix whose diagonal elements are the standard deviations of the Y_{ij}

under this model. Let

$$A_i^{1/2} = \begin{pmatrix} \{\phi V(\mu_{i1})\}^{1/2} & 0 & \cdots & 0 \\ 0 & \{\phi V(\mu_{i2})\}^{1/2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \{\phi V(\mu_{in_i})\}^{1/2} \end{pmatrix} \tag{3.8}$$

Let R_i be the $n_i \times n_i$ correlation matrix under one of the assumption above. Then we can write the covariance matrix for vector Y_{ij} as

$$\Sigma_i = \phi A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}. \tag{3.9}$$

We have the following statistical model for the mean vector and the covariance matrix of a data vector Y_i consisting of observations Y_{ij}, $j = 1, ..., n_i$ on unit i.

$$E(\boldsymbol{Y_i}) = \begin{pmatrix} f(x'_{i1}\boldsymbol{\beta}) \\ f(x'_{i2}\boldsymbol{\beta}) \\ \vdots \\ f(x'_{in_i}\boldsymbol{\beta}) \end{pmatrix} = f_i(\boldsymbol{\beta}), \qquad \text{var}(\boldsymbol{Y_i}) = \phi A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}. \tag{3.10}$$

3.2 Generalized Estimating Equations

The considerations in the last section allows for specification of a model for the mean and covariance of a data vector of the form (3.10). It is not possible to use the principle of maximum likelihood to develop a framework for estimation and testing. Although we don't have a basis for maximum likelihood, we are trying to emulate situations where there is such a basis:

- The normal case with mean model, the model is

  $$E(Y_i) = X_i\boldsymbol{\beta}, \qquad \text{var}(Y_i) = \Sigma_i \tag{3.11}$$

  for suitable choice of covariance matrix of Σ_i depending on a vector of parameters ω. Assuming that the Y_i follow a multivariate normal, the estimator for $\boldsymbol{\beta}$

  $$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{m} X_i'\hat{\Sigma}_i^{-1}X_i)^{-1} \sum_{i=1}^{m} X_i'\hat{\Sigma}_i^{-1}Y_i, \tag{3.12}$$

where $\hat{\Sigma}_i$ is the covariance matrix with the estimator for ω. It can be shown that it is possible to rewrite (3.12) in the form

$$\sum_{i=1}^{m} X_i' \hat{\Sigma}_i^{-1} (Y_i - X_i \hat{\beta}) = 0. \tag{3.13}$$

- In the case of ordinary generalized linear models, $\hat{\beta}$ is the solution to

$$\sum_{j=1}^{n} \frac{1}{V\{f(x_j'\beta)\}} \{Y_j - f(x_j'\beta)\} f'(x_j'\beta) x_j = 0, \tag{3.14}$$

  where $f' = \frac{d}{d\mu} f(\mu)$, the derivative of f with respect to its argument.

- Comparing (3.13) and (3.14), we see that there is a similar theme; the equations are linear function of deviation of observations from their assumed mean, weighted in accordance with their covariance and variance.

From these observations, a natural way to fit the model (3.10) is suggested: solve an estimating equation consisting of p equations for $\beta_{(p \times 1)}$ that

- is a linear function of deviations

$$Y_i - f_i(\beta). \tag{3.15}$$

- Weight these deviations in the same way as in (3.13) and (3.14), using the inverse of the assumed covariance matrix Σ_i of a data vector with an estimator for unknown parameters ω in the working correlation matrix. Even if there is a scale parameter we really need use only the inverse of A_i in (3.14). As in (3.14), Σ_i and A_i will also depend on β through the variance function $V\{f(x_{ij}'\beta)\}$

These results lead to consideration of the following equation to be solved for β

$$\sum_{i=1}^{m} \mathcal{D}_i' \hat{A}_i^{-1} \{Y_i - f_i(\hat{\beta})\} = 0, \tag{3.16}$$

where $\mathcal{D}_i$ is the $(n_i \times p)$ matrix whose element is the derivative of $f(x_{ij}'\beta)$ with respect to the sth element of β, and $\hat{A}_i$ is the matrix A_i with an estimator ω. An equation like (3.16) to be solved to estimate a parameter β in a mean response model is referred to as a **generalized estimating equation**.

The sampling distribution and hypothesis testing is the same as in a generalized linear model.

3.3 Robust Estimator for Sampling Covariance

It is important to recognize that GEE fitting method for estimating the parameters (mean and variance) is not a maximum likelihood method, rather, it was arrived at from an ad hoc perspective. As a result, it is not possible to derive quantities like AIC and BIC to compare different "working" correlation matrices to determine which assumption is not suitable. It is sensible to be concerned that the validity of inferences on $\boldsymbol{\beta}$ because calculation of approximate confidence intervals and tests may be compromised if the correlations assumptions on correlation is incorrect. One way to solve the problem is to modify the covariance matrix $\widehat{V_\beta}$ to allow the possibility that the choice of $R(\alpha)$ used in the model is incorrect. The modified version of $\widehat{V_\beta}$ is

$$\widehat{V}_\beta^R = \left(\sum_{i=1}^k \widehat{\mathbf{D}}_i' \widehat{V}^{-1} \widehat{\mathbf{D}}_i\right)^{-1} \left(\sum_{i=1}^k \widehat{\mathbf{D}}_i' \widehat{V}_i^{-1} \widehat{\mathbf{S}}_i \widehat{V}_i^{-1} \widehat{\mathbf{D}}_i\right) \left(\sum_{i=1}^k \widehat{\mathbf{D}}_i' \widehat{V}_i^{-1} \widehat{\mathbf{D}}_i\right)^{-1}, \tag{3.17}$$

where

$$\widehat{\mathbf{S}}_i = \{Y_i - \mu_i(\hat{\boldsymbol{\beta}})\}\{Y_i - \mu_i(\hat{\boldsymbol{\beta}})\}'. \tag{3.18}$$

$\widehat{V}_\beta^R$ is a robust estimate and will provide a reliable estimate of the true sampling covariance matrix of $\hat{\boldsymbol{\beta}}$ even if the chosen R_i is incorrect. $\widehat{V}_\beta$ is referred as the model-based covariance estimate.

3.4 GEE with Poisson data

For data in the form of counts, we have noted that a sensible probability model is the poisson distribution. This model dictates that variance is equal to the mean; moreover, any sensible representation of the mean ought to be such that the mean is forced to be positive.

(i) Mean: For regression modeling, we wish to represent the mean for Y_j as a function of the covariates X_j. However, this representation should ensure the mean can only be positive. A model that would accomplish this is

$$E(Y_j) = \exp(\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + ... + \beta_k x_{jk}) = \exp(\boldsymbol{x_j}'\boldsymbol{\beta}). \tag{3.19}$$

The positive requirement is enforced by writing the mean as the exponential of the linear function of $\boldsymbol{x_j}'\boldsymbol{\beta}$. The logarithm of the mean response is being modeled as a linear function of covariates and regression parameters,

$$\log E(Y_j) = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + ... + \beta_k x_{jk} = x_j'\boldsymbol{\beta}. \tag{3.20}$$

Therefore this is called log-linear model.

(ii) probability distribution: The Y_j are assumed to arise at each setting x_j from a poisson distribution.

(iii) Variance: Under the poisson assumption and the mean model, the variance of Y_j is given by

$$\mathrm{V}(Y_j) = E(Y_j) = \exp(x_j'\boldsymbol{\beta}). \tag{3.21}$$

The *PROC GENMOD* procedure in SAS uses the generalized estimating equations to extend the generalized model for repeated measures. We can use *PROC GENMOD* to perform a Poisson regression analysis of count data with log link function.

Chapter 4

Goodness of Fit and Model Selection

4.1 Approach

There seem to be few model-selection criteria available in GEE. The well-known Akaike Information Criterion (AIC) cannot be directly applied since AIC is based on maximum likelihood estimation while GEE is not likelihood based. Residuals from GEE regression models should be checked for the presence of outlier values that may seriously affect the results. Measures that test for the influence of a panel or case in the regression equation are extensions of those used in generalized linear models and are similar to those used in OLS regression. DFBETA measures the change in the fitted coefficient vector when a case is removed and is a measure of influence that can be used to analyze outliers and determine whether there are issues in the data that need further investigation. A visual test of the fitted GEE model that has been estimated is to plot residual versus fitted for each individual panel. In visually testing the residuals, a researcher should look for patterns that suggest a random distribution of residuals; they should not be clustered around certain values. For example, if a researcher saw that there were a large number of residuals with small negative values and a small number of high positive values, then different distribution and correlation structures should be examined. Another example would be the case in which there are changes in the pattern of the residuals across the time periods; this could indicate that they depend on the panel identifier and/or on the time identifier, and a different correlation structure should then be specified. SAS programs that fit GEEs provide users with the functionality to display residuals and DFBETA diagnostic statistics for observations in the data set.

4.2 Goodness of Fit Method

In this case, graphical and numerical methods for model assessment based on the cumulative sums of residual over some related aggregates of residuals will be used to check model fitting. Let Y_{ij} and X_{ij} be the same as previous section. The marginal mean of the response $\mu_{ij} = E(y_{ij})$ is assumed to depend on the covariate vector by $g(\mu_{ij} = X'_{ij}\beta)$.
Define the vector of residuals for the ith cluster as

$e_i = (e_{i1}, ..., e_{in_i})' = (y_{i1} - \hat{\mu}_{i1}, ..., y_{in_i} - \hat{\mu}_{in_i})'$.
Then we use the cumulative sum of the residuals with respect to x_{ip} to check the fit.

$$W_p(x) = \frac{1}{\sqrt{K}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I(x_{ijp} \leq x) e_{ij}, \tag{4.1}$$

where x_{ijp} is the pth component of X_{ij}. The null distribution of $W_p(x)$ can be approximated by the conditional distribution of

$$W_p(x) = \frac{1}{\sqrt{K}} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} I(x_{ijp} \leq x) e_{ij} + \eta'(x, \hat{\beta}) I_0^{-1} \hat{D}_i' \hat{V}_i^{-1} e_i \right) Z_i, \tag{4.2}$$

where $\hat{D}_i'$ and $\hat{V}_i^{-1}$ are the same as in chapter 3,

$$\eta(x, \beta) = -\sum_{i=1}^{m} \sum_{j=1}^{n_i} I(x_{ijp} \leq x) \frac{\partial \mu_{ij}}{\partial \beta}, \tag{4.3}$$

$$I_0 = \sum_{i=1}^{m} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i' \tag{4.4}$$

and the Z_i's are independent $N(0, 1)$ random variables.

4.3 Model Selection

4.3.1 Group

According to each subject's baseline characteristics, subjects were divided into two groups Asthma/Non-Asthma, and Atopic/Non-Atopic. For this analysis, we will have in total eight subgroup analyses: Asthmatic, Non-asthmatic, Atopic, Non-Atopic, Asthmatic and Non-Atopic, Non-Asthmatic and Atopic and Non-Asthmatic and Non-Atopic. Our response variables are LRT, URT and the sum of LRT and URT. We have 95 explanatory variables in total, as discussed in section 2.

4.3.2 Stepwise Regression to Detect Main Factors

Step 1: Fit a model with each linear predictor, the predictor with the smallest p-value is kept in the model, i.e. p1 (Detail SAS code in Appendix C)

Step 2: Fit a model with the predictor p1 in step 1 and plus each of the other predictors one by one. The model contains p1 and the other predictor with the smallest p-value will be continued for the next step. This is continued until p-value is greater than 0.15.

Step3: Check whether the p-value for the previous predictors in model are greater than 0.15 when adding new predictor.

4.3.3 Stepwise Regression to Detect Interaction Among Main Factors

Once for each group we find the model with the main predictors, we try to fit an interaction between the main factors in a stepwise fashion too. The macro (Appendix D)creates all the possible interaction terms for all the main factors remained in the model. The model building continues until the p-value for the addition of each interaction not in the model is greater than 0.15.

Chapter 5

Modeling Example

5.1 Modeling Example: Model Based vs. Robust Based Methods

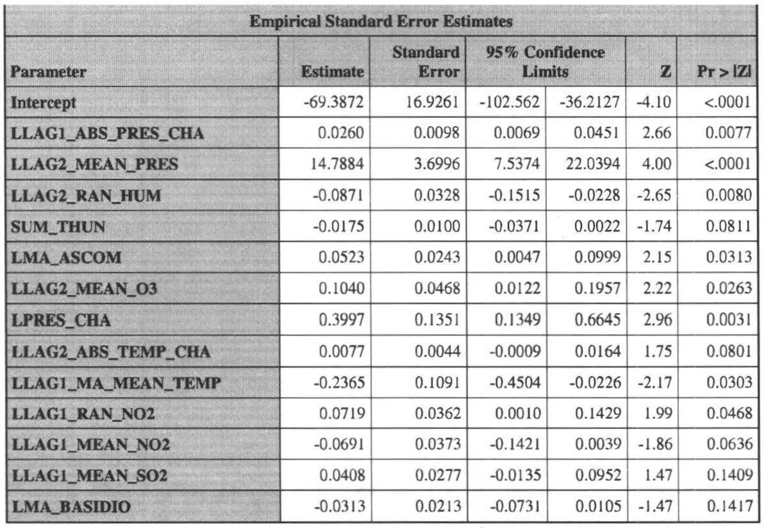
From Figure 5.1, since the estimates from model based and empirical based (3.17) show strong agreement. LLAG2-MEAN-PRES is significant in both models, LPRES-CHA and LLAG2-RAN-HUM are partially significant in both models as well. Therefore our specified working correlation matrix is a good choice.

5.2 Goodness of Fit Example

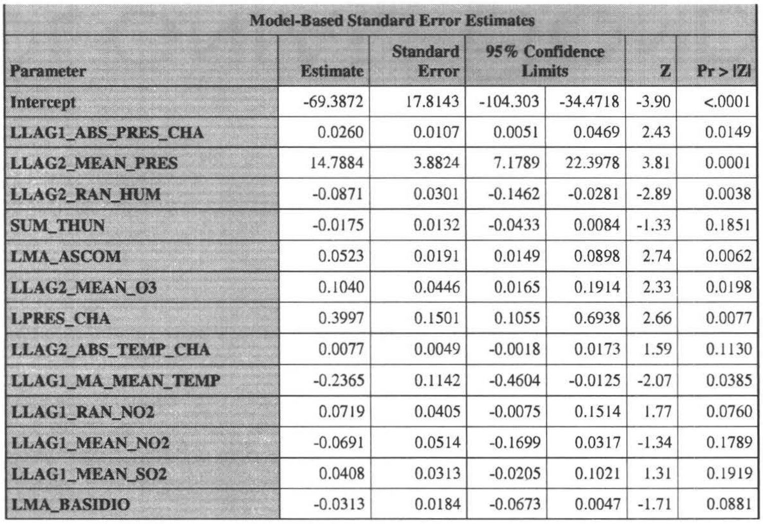
The solid line in Figure 5.2 shows the cumulative sum of residuals from the Asthmatic group with respect to logarithm of two-day lag average pressure (LLAG2-MEAN-PRES). For any value x on the horizontal axis, the solid line represents the cumulative sum of the residuals for all values of LLAGE2-MEAN-PRES less than or equal to x. Like the raw residuals, cumulative residuals will be centered at zero if the model fit is correct. The motivation for considering cumulative sums of residuals is that the asymptotic distribution can be determined. Under the null hypothesis of a correct model fit, they can be approximated as a zero mean Gaussian process with a covariance structure determined by the particular type of regression model (4.2). Realizations of the Gaussian process can be simulated by computer and compared with the observed process to assess whether the observed residual process represents anything beyond random variation. The light dashed lines in Figure 4.1 are the first 20 realizations of 10,000 simulated paths of the cumulative residual process under the null hypothesis of a correct model fit. Most of the paths tend to be closer to and intersect the horizontal axis more than the observed residuals. The maximum absolute value of the observed cumulative residuals is 25. Of the 10,000 realizations under the null hypothesis, the p-value for a Kolmogorov-type supremum test is 0.02. This tells us that the functional form of LLAG2-MEAN-PRES is adequate for the model.

In figure 5.3, The p-value of 0.1010 suggests that a more adequate model may be possible. The observed cumulative residuals represented by the heavy lines, seem atypical of the simulated curves, represented by the light lines, reinforcing the conclusion that a more appropriate functional form for pressure is possible.

Figure 5.1: *Comparison between model based and empirical based methods*



Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-69.3872	16.9261	-102.562	-36.2127	-4.10	<.0001
LLAG1_ABS_PRES_CHA	0.0260	0.0098	0.0069	0.0451	2.66	0.0077
LLAG2_MEAN_PRES	14.7884	3.6996	7.5374	22.0394	4.00	<.0001
LLAG2_RAN_HUM	-0.0871	0.0328	-0.1515	-0.0228	-2.65	0.0080
SUM_THUN	-0.0175	0.0100	-0.0371	0.0022	-1.74	0.0811
LMA_ASCOM	0.0523	0.0243	0.0047	0.0999	2.15	0.0313
LLAG2_MEAN_O3	0.1040	0.0468	0.0122	0.1957	2.22	0.0263
LPRES_CHA	0.3997	0.1351	0.1349	0.6645	2.96	0.0031
LLAG2_ABS_TEMP_CHA	0.0077	0.0044	-0.0009	0.0164	1.75	0.0801
LLAG1_MA_MEAN_TEMP	-0.2365	0.1091	-0.4504	-0.0226	-2.17	0.0303
LLAG1_RAN_NO2	0.0719	0.0362	0.0010	0.1429	1.99	0.0468
LLAG1_MEAN_NO2	-0.0691	0.0373	-0.1421	0.0039	-1.86	0.0636
LLAG1_MEAN_SO2	0.0408	0.0277	-0.0135	0.0952	1.47	0.1409
LMA_BASIDIO	-0.0313	0.0213	-0.0731	0.0105	-1.47	0.1417



Model-Based Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-69.3872	17.8143	-104.303	-34.4718	-3.90	<.0001
LLAG1_ABS_PRES_CHA	0.0260	0.0107	0.0051	0.0469	2.43	0.0149
LLAG2_MEAN_PRES	14.7884	3.8824	7.1789	22.3978	3.81	0.0001
LLAG2_RAN_HUM	-0.0871	0.0301	-0.1462	-0.0281	-2.89	0.0038
SUM_THUN	-0.0175	0.0132	-0.0433	0.0084	-1.33	0.1851
LMA_ASCOM	0.0523	0.0191	0.0149	0.0898	2.74	0.0062
LLAG2_MEAN_O3	0.1040	0.0446	0.0165	0.1914	2.33	0.0198
LPRES_CHA	0.3997	0.1501	0.1055	0.6938	2.66	0.0077
LLAG2_ABS_TEMP_CHA	0.0077	0.0049	-0.0018	0.0173	1.59	0.1130
LLAG1_MA_MEAN_TEMP	-0.2365	0.1142	-0.4604	-0.0125	-2.07	0.0385
LLAG1_RAN_NO2	0.0719	0.0405	-0.0075	0.1514	1.77	0.0760
LLAG1_MEAN_NO2	-0.0691	0.0514	-0.1699	0.0317	-1.34	0.1789
LLAG1_MEAN_SO2	0.0408	0.0313	-0.0205	0.1021	1.31	0.1919
LMA_BASIDIO	-0.0313	0.0184	-0.0673	0.0047	-1.71	0.0881

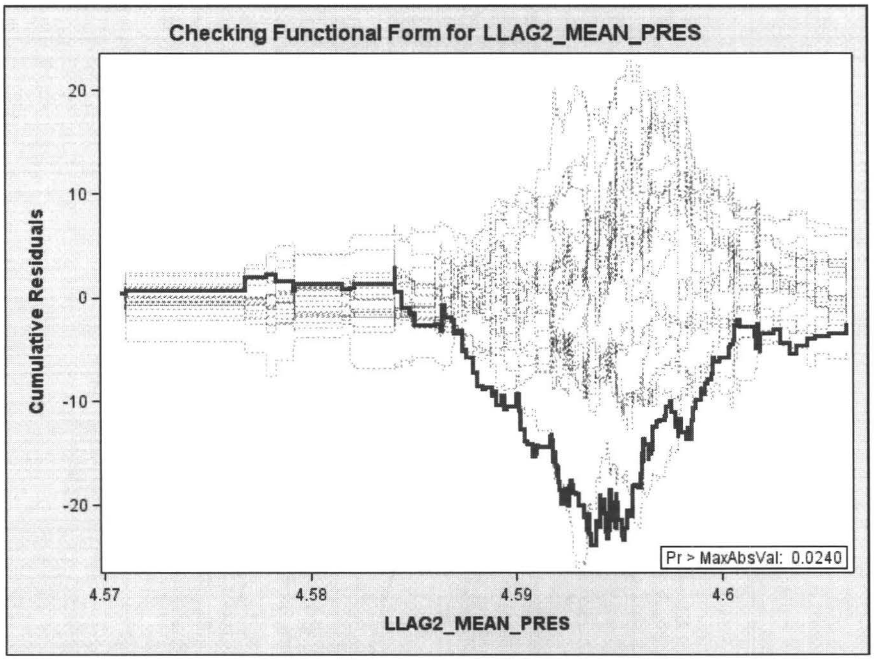


Figure 5.2: *Cumulative Residual Plot for Asthmatic Group*

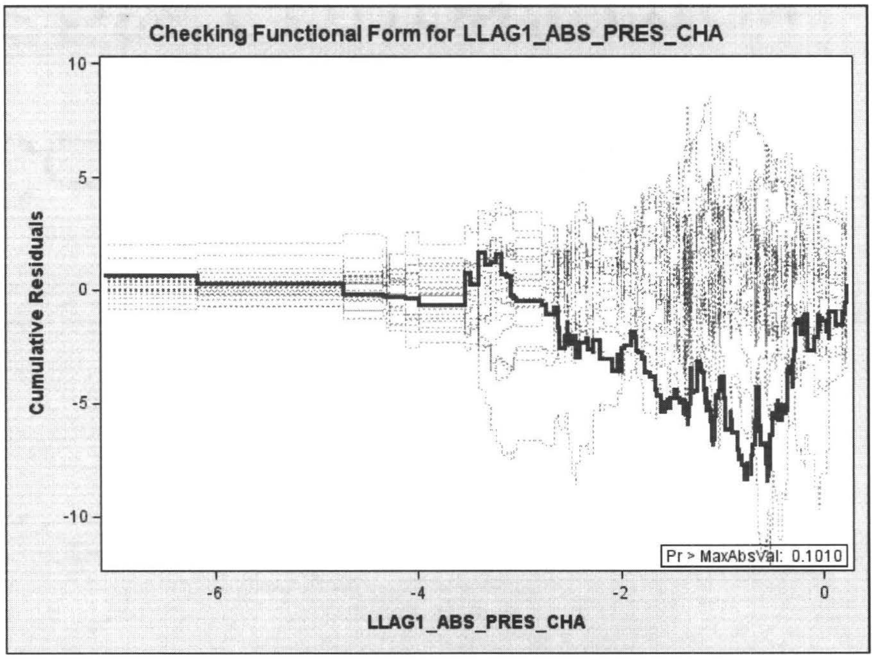


Figure 5.3: *Cumulative Residual Plot for Non-Atopic Group*

Chapter 6

Results

In this chapter the results of eight group analysis for this project are presented and discussed. The models are presented in the table format along with the predictors. The significant predictors and possible interaction between significant predictors are displayed.

Table 6.1: *Significant predictors for different groups*:

Group	Variable	P-value	Variable	P-value
Asthmatic	LLAG2-MEAN-PRES	0.0022(**)	LRAN-NO	0.0083(**)
	LLAG2-RAN-HUM	0.0306	SUM-THUN	0.0311
	LMA-ASCOM	0.0381	LLAG1-ABS-PRES-CHA	0.0398
Atopic	LLAG2-MEAN-PRES	0.0004(***)	LLAG1-MEAN-NO2	0.0102
	LLAG2-RAN-HUM	0.0138	LLAG1-MEAN-SO2	0.0175
	LAG2-RAN-NOX	0.0295	LMA-MEAN-O3	0.0433
Non-Asthmatic	LLAG1-MEAN-NO2	0.0001(***)	LRAN-TEMP	0.0168
	LMEAN-SO2	0.0264	LABS-TEMP-CHA	0.0356
	LABS-TEMP-CHA * LRAN-TEMP	0.0359	LRAN-CO	0.0369
Non-Atopic	LMA-MEAN-CO	< 0.0001 (***)	AGE GROUP	0.0099(**)
	LRAN-NO	0.0047(**)	LLAG1-ABS-PRES-CHA	0.0305
	LLAG2-MA-WEEDS * AGE GROUP	0.0404	LABS-TEMP-CHA	0.0410
Asthmatic and Atopic	LLAG2-MEAN-PRES	0.0015(**)	LLAG1-RAN-TEMP	0.0060(**)
	LAG2-SUM-THUN	0.0436		
Asthmatic and Non-Atopic	LLAG1-MA-POLLEN	0.0012(**)	LLAG1-RAN-NO2	0.0064(**)
	LLAG2-MA-MEAN-CO	0.0117	LRAN-HUM	0.0206
	LRAN-PRES	0.0313	LLAG1-MA-WEEDS*LRAN-HUM	0.041
Non-Asthmatic and Atopic	LMEAN-SO2	0.0378		
Non-Asthmatic and Non-Atopic	LLAG1-MA-WEEDS	0.0002(***)	LLAG1-MEAN-NO2	0.0005(***)
	LLAG1-MEAN-NO2* LLAG1-MA-WEEDS	0.0009(***)		

Note:
**: significant at a 1% level of significance
***: significant at a 0.1% level of significance

It should be noted that the above table only shows the predictors that are significant at a 5% level of significance for different groups. The log transformation of two-day lag of mean pressure is significant in both asthmatic group and atopic group at a 0.1% level of significance. By examing the p-value, both weather parameter and air pollution parameter are significant predictors for asthmatic children. There seems no strong evidence that there are any interaction between weather, air pollution and air pollen are significant associated with asthmatic children. There are two groups that don't have any interaction term in the final model: Asthmatic and Atopic, Non-Asthmatic and Atopic.

The detail results of the 8 groups are shown below: For each group, the first part of the table shows the final model without interactions. The second part of tables shows the final model with interaction terms. The p-values that are < 0.01 are highlighted.

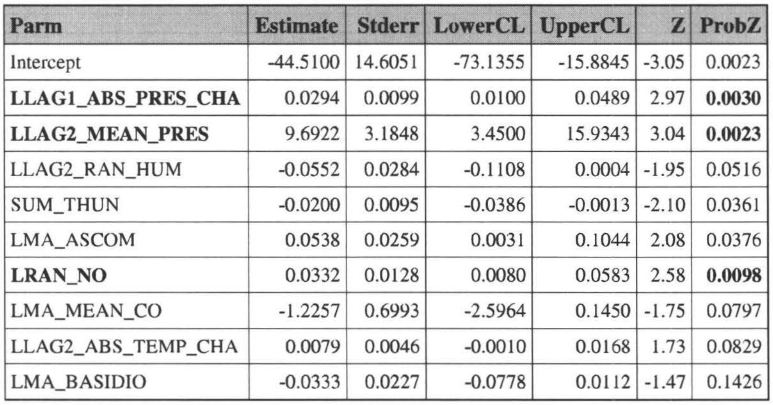
Results for Asthmatic group (Figure 6.1): 9 predictors are kept in the final model with one interaction term. The significant predictors are log transformation of two-day lag mean pressure and log transformation of range in Nitric Oxide. Both of the positive coefficient signs mean that an increase in the predictors will resolve an increase in lrt count. The significant interaction term is log transformation of one-day lag of absolute pressure change and log transformation of two day lag range of humidity. This means that as LLAG1-ABS-PRES-CHA increases, the effect of humidity on lrt count gets greater.

Results for Non-Asthmatic group (Figure 6.2): 9 predictors are in the final model with 3 interaction terms. The significant predictors are log transformation of one-day lag mean of Nitrogen Dioxide and range of temperature. Both of the negative coefficient signs mean that an increase in the predictors will result an decrease in lrt count on average population.

Results for the other groups can be found in the Appendix. From table 6.1, we conclude that the fact that we have so many predictors in the model will lead to too many significant results inflated type I error. Overall, the significant factor that is associated with LRT scores for children with asthma is the two-day lag daily average changes in air pressure. On average an increase in air pressure will result in an increase in children's LRT scores. The interaction terms that remained in the final model show some degree of significance but without strong evidence to support it. Children in the case groups are more sensitive to meteorological factors, allergens, and air pollution than the children in control groups.

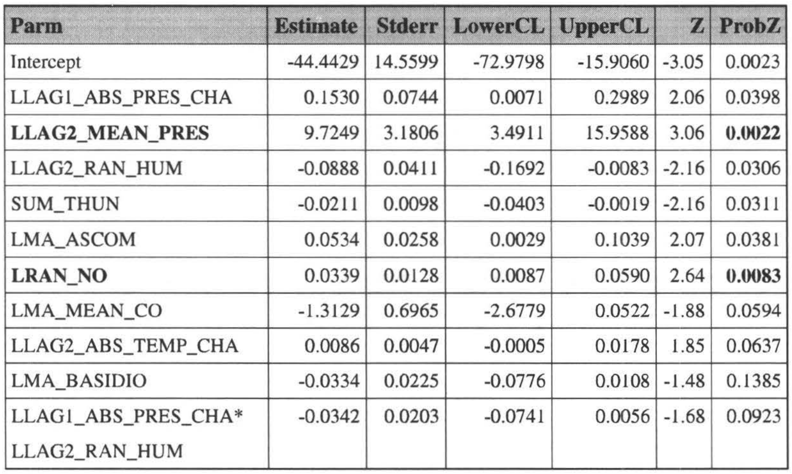
Figure 6.1: *Results for Asthmatic Group*

Final model without interaction for Asthmatic group



Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-44.5100	14.6051	-73.1355	-15.8845	-3.05	0.0023
LLAG1_ABS_PRES_CHA	0.0294	0.0099	0.0100	0.0489	2.97	**0.0030**
LLAG2_MEAN_PRES	9.6922	3.1848	3.4500	15.9343	3.04	**0.0023**
LLAG2_RAN_HUM	-0.0552	0.0284	-0.1108	0.0004	-1.95	0.0516
SUM_THUN	-0.0200	0.0095	-0.0386	-0.0013	-2.10	0.0361
LMA_ASCOM	0.0538	0.0259	0.0031	0.1044	2.08	0.0376
LRAN_NO	0.0332	0.0128	0.0080	0.0583	2.58	**0.0098**
LMA_MEAN_CO	-1.2257	0.6993	-2.5964	0.1450	-1.75	0.0797
LLAG2_ABS_TEMP_CHA	0.0079	0.0046	-0.0010	0.0168	1.73	0.0829
LMA_BASIDIO	-0.0333	0.0227	-0.0778	0.0112	-1.47	0.1426

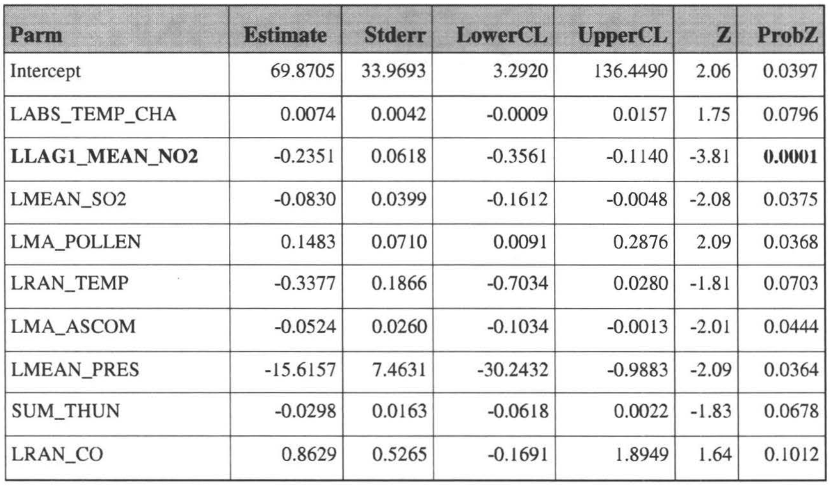
Final model with interaction for Asthmatic group



Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-44.4429	14.5599	-72.9798	-15.9060	-3.05	0.0023
LLAG1_ABS_PRES_CHA	0.1530	0.0744	0.0071	0.2989	2.06	0.0398
LLAG2_MEAN_PRES	9.7249	3.1806	3.4911	15.9588	3.06	**0.0022**
LLAG2_RAN_HUM	-0.0888	0.0411	-0.1692	-0.0083	-2.16	0.0306
SUM_THUN	-0.0211	0.0098	-0.0403	-0.0019	-2.16	0.0311
LMA_ASCOM	0.0534	0.0258	0.0029	0.1039	2.07	0.0381
LRAN_NO	0.0339	0.0128	0.0087	0.0590	2.64	**0.0083**
LMA_MEAN_CO	-1.3129	0.6965	-2.6779	0.0522	-1.88	0.0594
LLAG2_ABS_TEMP_CHA	0.0086	0.0047	-0.0005	0.0178	1.85	0.0637
LMA_BASIDIO	-0.0334	0.0225	-0.0776	0.0108	-1.48	0.1385
LLAG1_ABS_PRES_CHA* LLAG2_RAN_HUM	-0.0342	0.0203	-0.0741	0.0056	-1.68	0.0923

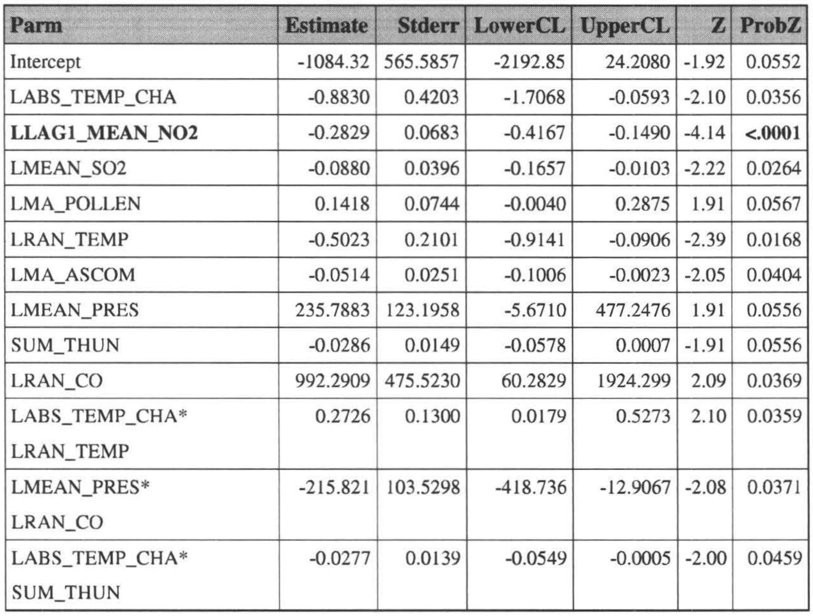
Figure 6.2: *Results for Non-Asthmatic Group*

Final model with interaction for Non-Asthmatic group



Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	69.8705	33.9693	3.2920	136.4490	2.06	0.0397
LABS_TEMP_CHA	0.0074	0.0042	-0.0009	0.0157	1.75	0.0796
LLAG1_MEAN_NO2	-0.2351	0.0618	-0.3561	-0.1140	-3.81	**0.0001**
LMEAN_SO2	-0.0830	0.0399	-0.1612	-0.0048	-2.08	0.0375
LMA_POLLEN	0.1483	0.0710	0.0091	0.2876	2.09	0.0368
LRAN_TEMP	-0.3377	0.1866	-0.7034	0.0280	-1.81	0.0703
LMA_ASCOM	-0.0524	0.0260	-0.1034	-0.0013	-2.01	0.0444
LMEAN_PRES	-15.6157	7.4631	-30.2432	-0.9883	-2.09	0.0364
SUM_THUN	-0.0298	0.0163	-0.0618	0.0022	-1.83	0.0678
LRAN_CO	0.8629	0.5265	-0.1691	1.8949	1.64	0.1012

Final model without interaction for Non-Asthmatic group



Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-1084.32	565.5857	-2192.85	24.2080	-1.92	0.0552
LABS_TEMP_CHA	-0.8830	0.4203	-1.7068	-0.0593	-2.10	0.0356
LLAG1_MEAN_NO2	-0.2829	0.0683	-0.4167	-0.1490	-4.14	**<.0001**
LMEAN_SO2	-0.0880	0.0396	-0.1657	-0.0103	-2.22	0.0264
LMA_POLLEN	0.1418	0.0744	-0.0040	0.2875	1.91	0.0567
LRAN_TEMP	-0.5023	0.2101	-0.9141	-0.0906	-2.39	0.0168
LMA_ASCOM	-0.0514	0.0251	-0.1006	-0.0023	-2.05	0.0404
LMEAN_PRES	235.7883	123.1958	-5.6710	477.2476	1.91	0.0556
SUM_THUN	-0.0286	0.0149	-0.0578	0.0007	-1.91	0.0556
LRAN_CO	992.2909	475.5230	60.2829	1924.299	2.09	0.0369
LABS_TEMP_CHA* LRAN_TEMP	0.2726	0.1300	0.0179	0.5273	2.10	0.0359
LMEAN_PRES* LRAN_CO	-215.821	103.5298	-418.736	-12.9067	-2.08	0.0371
LABS_TEMP_CHA* SUM_THUN	-0.0277	0.0139	-0.0549	-0.0005	-2.00	0.0459

Chapter 7

Discussion and Future Work

In this project, we have shown the association of meteorological factors, allergens and air pollution on asthma in children and their possible interaction effect. Even though the asthmatic and atopic groups show some degree of agreement, we can include a seasonal component, virus date, September school return and holiday periods in the model to check whether these factors are associated with children's LRT.

The limitation of this results is that in fact we have so many predictors in the model that will lead to too many significant results which inflated type I error.

In this analysis , we use SAS *proc genmod*, which we are discussing from the population-averaged specific. The rationale is that interest focuses on what happens on average in a population. But on the alternative, if we are interested in individual trajectories, we are referring to the statistical model as generalized linear mixed model. This can be analyzed in SAS using *proc glimmix.*

As we mentioned, a common issue with longitudinal data, particularly when the units are humans, is that some data may be missing. The obvious consequence is that the resulting data may not be balanced. Correcting the problem is difficult, when the missingness is a consequence of something we can't observe. If nonignorable nonresponse is suspected, it may not be possible to obtain reliable inferences without making assumptions.

Chapter 8

Appendix

8.1 Partial SAS Codes for Importing data

```
goptions reset=global;
/*macro print to print out SAS statement generated by macro execution*/;
options symbolgen mprint;
/*Import excel sheets for weather data*/
%macro weather(month=,year=);
%let start=&month&year;
%let s=sum;
%let t=thun;
PROC IMPORT DBMS=EXCEL OUT= work.&start
      DATAFILE="climate dataset\&start"  REPLACE ;
      RANGE="A10:Z754";
     GETNAMES=YES;        /* Use the first row of data as column names    */
     SCANTEXT=YES;    /* Scan all rows of data for the largest size  */
     USEDATE=YES;     /* Use DATE format for date/time columns    */
     SCANTIME=YES;    /* Scan and identify time columns           */
     DBSASLABEL=NONE;     /* Leave SAS label names to be nulls        */
RUN;

/*for each hour, see if there is an occurrence of thunderstorm*/
DATA &start;
 SET &start;
 where date_time <>.;
 NUM_THUN= 0;
   IF Weather___='Thunderstorms'  THEN DO;
        NUM_THUN=1;
```

```
  END;
RUN;

/* For each day,calculate the number of hours thunder storms
occurs*/

proc means data=&start noprint nway missing;
  class day;
  var  NUM_THUN;
  OUTPUT OUT=&t&start
               SUM=SUM_THUN;
RUN;

 /* calculate average  and range of TEMP, PRES , HUM  */
proc means data=&start NOPRINT NWAY;
 class  Day Date_time;
 VAR Temp___C_  Stn_Press__kPa_  Rel_Hum____  ;
 OUTPUT OUT=&start&s
           MEAN = MEAN_TEMP MEAN_PRES MEAN_HUM
          RANGE=RAN_TEMP RAN_PRES RAN_HUM;
run;
** merge thunderstorm variable with the rest weather data**;
data &s&start;
   merge  &start&s &t&start;
   drop  _TYPE_ _FREQ_ ;
RUN;

%mend weather;
quit;
** The reason to do it seperately is because the data come    **
** in month and each spread sheet contains extra information **
** that need to be deleted before we merge it.               **;
%weather (month=jul,year=03);  ** call macro to import all the dataset**;
%weather (month=aug, year=03);
%weather (month=sep, year=03);
%weather (month=oct, year=03):
%weather (month=nov, year=03);
%weather (month=dec, year=03);
%weather (month=jan, year=04);
%weather (month=feb, year=04);
%weather (month=mar, year=04);
%weather (month=apr, year=04):
%weather (month=may, year=04);
%weather (month=jun, year=04);
```

```
%weather (month=jul,year=04);
%weather (month=aug, year=04);
%weather (month=sep, year=04);
%weather (month=oct, year=04):
%weather (month=nov, year=04);
%weather (month=dec, year=04);

** merge all the dataset into one big dataset**;
DATA weather_sum;
  SET sumjul03 sumaug03 sumsep03 sumoct03 sumnov03
  sumdec03 sumjan04 sumfeb04 summar04 sumapr04 summay04 sumjun04
  sumjul04 sumaug04 sumsep04 sumoct04 sumnov04 sumdec04;
  TEMP_CHA=MEAN_TEMP-LAG(MEAN_TEMP); ** daily mean temp change**
  PRES_CHA=MEAN_PRES-LAG(MEAN_PRES);
  ABS_TEMP_CHA=ABS(TEMP_CHA);        ** absolute daily temp change**
  ABS_PRES_CHA=ABS(PRES_CHA);
RUN;
** Sort according to date time**;
proc sort data=weather_sum out=w_sum;
 by Date_Time;
run;
```

8.2 SAS Codes for Calculating moving average

```
**  This code is used to remove variables with tread or seasonal      **
**  effect. In this project, we decode to use 5 day moving average.   **;

 DATA moving_ave;
 SET mean_table;
 mean_temp_lag24=lag(mean_temp); ** one day lag of mean temp**;
 mean_temp_lag48=lag2(mean_temp);
 mean_temp_lag72=lag3(mean_temp);
 mean_temp_lag96=lag4(mean_temp);
 if _N_  GE 5 THEN  TEMP_AVE=MEAN(OF mean_temp mean_temp_:);
 drop mean_temp_lag24 mean_temp_lag48
 mean_temp_lag72 mean_temp_lag96;
run;

data diff_ma;
 set moving_ave;
 ma_mean_temp=(mean_temp-temp_ave);
 drop temp_ave;
run;
```

8.3 Partial SAS codes for Proc genmod

```
/* Get subjects in ashtmatic and atopic group*/;
 data asthmatic_atopic;
 set sasdata.longitudinal_table;
 where group3=1 and group2=1;
 drop group1 group2 group3 urt char1 char2;
run;

 /* Start stepwise regression analysis*/
 /* step 1 regression analysis with lrt*/
data step;
 set  asthmatic_atopic(drop=lrt);
run;

proc sort data=step;
 by Date_Time id;
run;

/* transpose col variables to row variables*/
 proc transpose data=step out=sum (rename=(col1=x_variable));
 by Date_Time id ;
run;

data info;
 set asthmatic_atopic;
 keep Date_Time id lrt;
run;

proc sort data=info out=info1;
 by  Date_Time id;
run;

/* Merge weather, aeroallergen, and airpollution data with data
    contains subject's LRT score*/
 data final_reg2;
 set sum info1;
 merge sum info1;
 by Date_Time id;
run;

proc sort data=final_reg2;
 by _name_;
run;
```

```
**Call proc genmod**;
 proc genmod data=final_reg2;
 by _name_;
 class id;
 model lrt=x_variable/ dist=poisson link=log;
 repeated subject=id/type=ar(1)  modelse; ** model based**;
 ods output GEEEmpPEst=parameterEst; **put output estimates in a seperate data **
run;

data reg2_table1;
  set parameterEst;
  where Parm='x_variable'; ** Keep the variable of interest**;
  keep _name_ Estimate Probz;
run;

/* vars and corresponding p value sorted in increasing*/
proc sort data=reg2_table1;
 by Probz ;
run;
 data reg1;
 set reg2_table1(obs=1);
 where probz LE &p_value; ** call macro variable: p-value=0.15**;
run; /*

create global macro variable*/
proc sql;
select _name_ into: var1
from reg1; quit;
** First variable in the model**;
%put var1 is &var1;
```

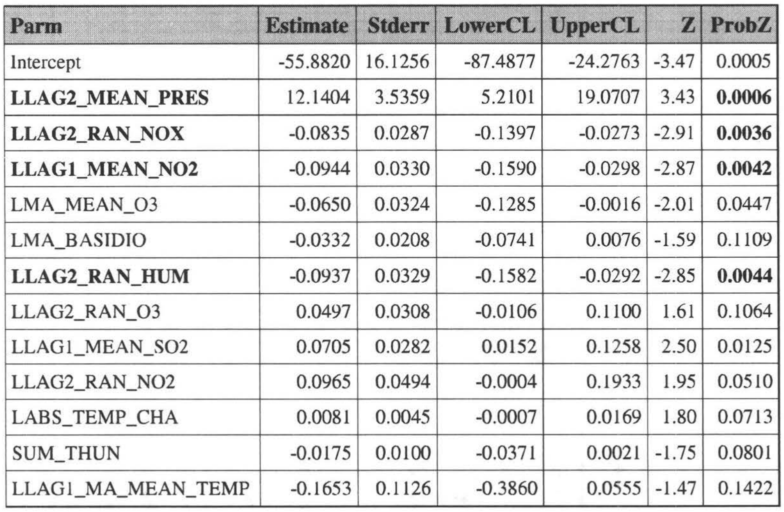
8.4 Macros used to produce interaction variables

```
** This macro is used once we have the final variables in the model. **
** We use this macro to create interaction terms for PROC GENMOD.    **;
%macro interact(
    data=_last_ ,    /* name of input dataset */
    out=&data,       /* name of output dataset */
    v= ,            /*  variable(s) */
    prefix = I_,     /* prefix for interaction variable names */
    names=,          /* or, a list of n*m names */
    center=          /* mean center first? */
    );
** check if the variable list is empty or not**;
%let abort = 0;
%if (%length(&v) = 0 ) %then %do;
        %put ERROR: INTERACT: V=  must be specified;
        %goto done;
        %end;
**Check if the dataset exist. Default is last used sas output data**;
%if %bquote(&data) = _last_ %then %let data = &syslast;
%if %bquote(&data) = _NULL_ %then %do;
    %put ERROR: There is no default input data set (_LAST_ is _NULL_);
    %goto DONE;
    %end;
** get standardize values**;
%if %length(&center) %then %do;
proc standard data=&data out=&data m=0;
    var &center;
    %end;

data &out;
    set &data;
** Local variables**;
    %local i j k w1 w2;

    %let k=0;
    %let i=1;
    %let w1 = %scan(&v, &i, %str( )); ** Get the first variable**;
    %do %while(&w1 ^= );

        %let j=%eval(&i+1);
        %let w2 = %scan(&v, &j, %str( )); ** Get the second variable**;

        %do %while(&w2 ^= );
            %* put i=&i j=&j;
            %let k=%eval(&k+1);
            %let name = %scan(&names, &k, %str( ));
            %if %length(&name) %then %do; ** If interaction names specified**;
                &name = &w1 * &w2;  ** Assigns interaction term**;
                %end;
            %else %do; ** If interaction names not specified, use default names**;
                &prefix.&i.&j = &w1 * &w2;
                %end;
            %let j=%eval(&j+1);
            %let w2 = %scan(&v, &j, %str( ));
            %end;

        %let i=%eval(&i+1);
        %let w1 = %scan(&v, &i, %str( ));
        %end;
run;

%done:
%if &abort %then %put ERROR: The INTERACT macro ended abnormally.;

%mend;
```

8.5 Results for other groups

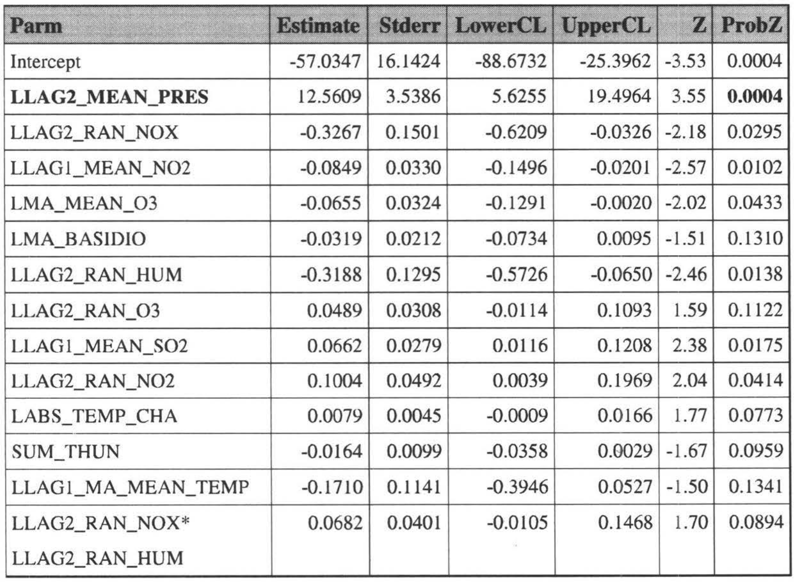
Figure 8.1: *Results for Atopic Group*

Final model without interaction for Atopic group



Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-55.8820	16.1256	-87.4877	-24.2763	-3.47	0.0005
LLAG2_MEAN_PRES	12.1404	3.5359	5.2101	19.0707	3.43	**0.0006**
LLAG2_RAN_NOX	-0.0835	0.0287	-0.1397	-0.0273	-2.91	**0.0036**
LLAG1_MEAN_NO2	-0.0944	0.0330	-0.1590	-0.0298	-2.87	**0.0042**
LMA_MEAN_O3	-0.0650	0.0324	-0.1285	-0.0016	-2.01	0.0447
LMA_BASIDIO	-0.0332	0.0208	-0.0741	0.0076	-1.59	0.1109
LLAG2_RAN_HUM	-0.0937	0.0329	-0.1582	-0.0292	-2.85	**0.0044**
LLAG2_RAN_O3	0.0497	0.0308	-0.0106	0.1100	1.61	0.1064
LLAG1_MEAN_SO2	0.0705	0.0282	0.0152	0.1258	2.50	0.0125
LLAG2_RAN_NO2	0.0965	0.0494	-0.0004	0.1933	1.95	0.0510
LABS_TEMP_CHA	0.0081	0.0045	-0.0007	0.0169	1.80	0.0713
SUM_THUN	-0.0175	0.0100	-0.0371	0.0021	-1.75	0.0801
LLAG1_MA_MEAN_TEMP	-0.1653	0.1126	-0.3860	0.0555	-1.47	0.1422

Final model with interaction for Atopic group



Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-57.0347	16.1424	-88.6732	-25.3962	-3.53	0.0004
LLAG2_MEAN_PRES	12.5609	3.5386	5.6255	19.4964	3.55	**0.0004**
LLAG2_RAN_NOX	-0.3267	0.1501	-0.6209	-0.0326	-2.18	0.0295
LLAG1_MEAN_NO2	-0.0849	0.0330	-0.1496	-0.0201	-2.57	0.0102
LMA_MEAN_O3	-0.0655	0.0324	-0.1291	-0.0020	-2.02	0.0433
LMA_BASIDIO	-0.0319	0.0212	-0.0734	0.0095	-1.51	0.1310
LLAG2_RAN_HUM	-0.3188	0.1295	-0.5726	-0.0650	-2.46	0.0138
LLAG2_RAN_O3	0.0489	0.0308	-0.0114	0.1093	1.59	0.1122
LLAG1_MEAN_SO2	0.0662	0.0279	0.0116	0.1208	2.38	0.0175
LLAG2_RAN_NO2	0.1004	0.0492	0.0039	0.1969	2.04	0.0414
LABS_TEMP_CHA	0.0079	0.0045	-0.0009	0.0166	1.77	0.0773
SUM_THUN	-0.0164	0.0099	-0.0358	0.0029	-1.67	0.0959
LLAG1_MA_MEAN_TEMP	-0.1710	0.1141	-0.3946	0.0527	-1.50	0.1341
LLAG2_RAN_NOX* LLAG2_RAN_HUM	0.0682	0.0401	-0.0105	0.1468	1.70	0.0894

Figure 8.2: *Results for Non-Atopic Group*

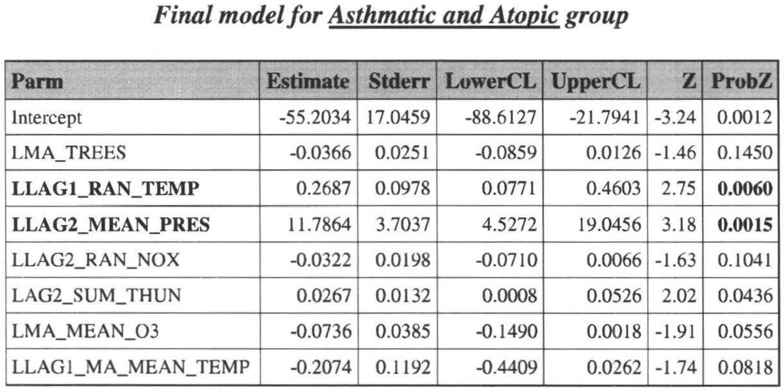
Final model without interaction for Non-Atopic group

Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	2.4269	1.4179	-0.3520	5.2059	1.71	0.0870
LMA_MEAN_CO	-5.5499	1.4404	-8.3731	-2.7267	-3.85	**0.0001**
LLAG2_MA_WEEDS	0.1642	0.0654	0.0361	0.2923	2.51	0.0120
LRAN_NO	0.1013	0.0354	0.0320	0.1706	2.87	**0.0042**
LABS_TEMP_CHA	0.0177	0.0085	0.0010	0.0344	2.08	0.0379
LMA_ASCOM	0.0697	0.0353	0.0006	0.1388	1.98	0.0481
LLAG1_ABS_PRES_CHA	0.0443	0.0209	0.0033	0.0853	2.12	0.0340
AGE GROUP	-0.6764	0.3773	-1.4158	0.0630	-1.79	0.0730
GENDER	0.7146	0.4422	-0.1521	1.5813	1.62	0.1061
LLAG2_MEAN_NO	0.0466	0.0264	-0.0051	0.0983	1.77	0.0772
LRAN_PRES	0.0551	0.0346	-0.0127	0.1229	1.59	0.1112

Final model with interaction for Non-Atopic group

Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	5.0496	1.7274	1.6639	8.4353	2.92	0.0035
LMA_MEAN_CO	-5.4856	1.4375	-8.3030	-2.6682	-3.82	**0.0001**
LLAG2_MA_WEEDS	-0.4391	0.2647	-0.9579	0.0798	-1.66	0.0972
LRAN_NO	0.1001	0.0354	0.0306	0.1695	2.82	**0.0047**
LABS_TEMP_CHA	0.0173	0.0085	0.0007	0.0339	2.04	0.0410
LMA_ASCOM	0.0694	0.0352	0.0004	0.1384	1.97	0.0488
LLAG1_ABS_PRES_CHA	0.0456	0.0211	0.0043	0.0868	2.16	0.0305
AGE GROUP	-2.9765	1.1545	-5.2392	-0.7138	-2.58	**0.0099**
GENDER	0.6964	0.4421	-0.1702	1.5630	1.58	0.1152
LLAG2_MEAN_NO	0.0470	0.0264	-0.0048	0.0988	1.78	0.0755
LRAN_PRES	0.0571	0.0346	-0.0107	0.1248	1.65	0.0987
LLAG2_MA_WEEDS* AGE GROUP	0.5226	0.2550	0.0229	1.0224	2.05	0.0404

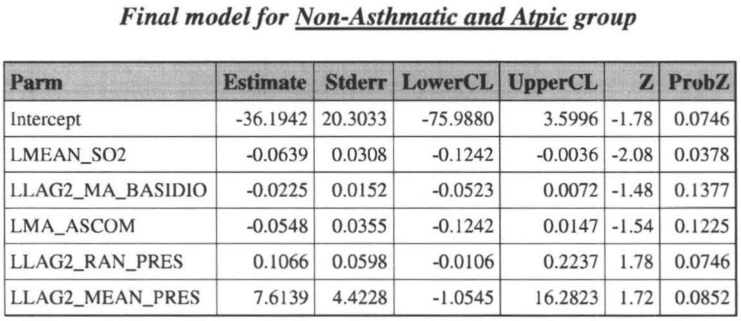
Figure 8.3: *Results for Asthmatic and Atopic Group*



Final model for Asthmatic and Atopic group

Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-55.2034	17.0459	-88.6127	-21.7941	-3.24	0.0012
LMA_TREES	-0.0366	0.0251	-0.0859	0.0126	-1.46	0.1450
LLAG1_RAN_TEMP	0.2687	0.0978	0.0771	0.4603	2.75	**0.0060**
LLAG2_MEAN_PRES	11.7864	3.7037	4.5272	19.0456	3.18	**0.0015**
LLAG2_RAN_NOX	-0.0322	0.0198	-0.0710	0.0066	-1.63	0.1041
LAG2_SUM_THUN	0.0267	0.0132	0.0008	0.0526	2.02	0.0436
LMA_MEAN_O3	-0.0736	0.0385	-0.1490	0.0018	-1.91	0.0556
LLAG1_MA_MEAN_TEMP	-0.2074	0.1192	-0.4409	0.0262	-1.74	0.0818

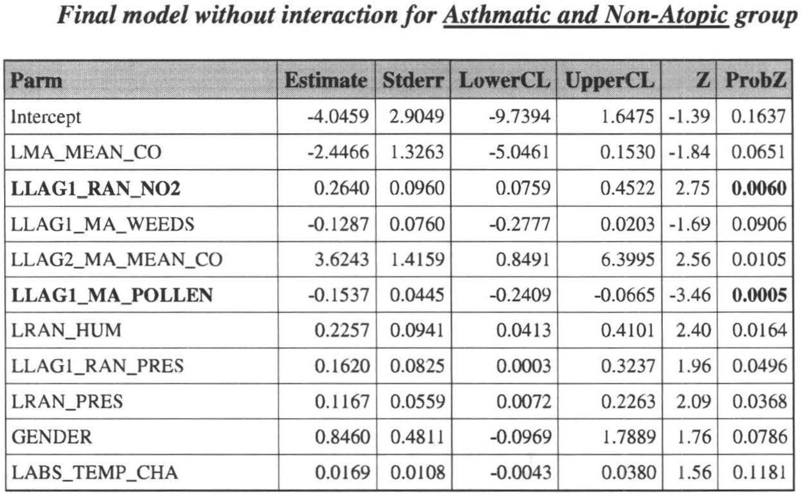
Figure 8.4: *Results for Non-Asthmatic and Atopic Group*



Final model for Non-Asthmatic and Atpic group

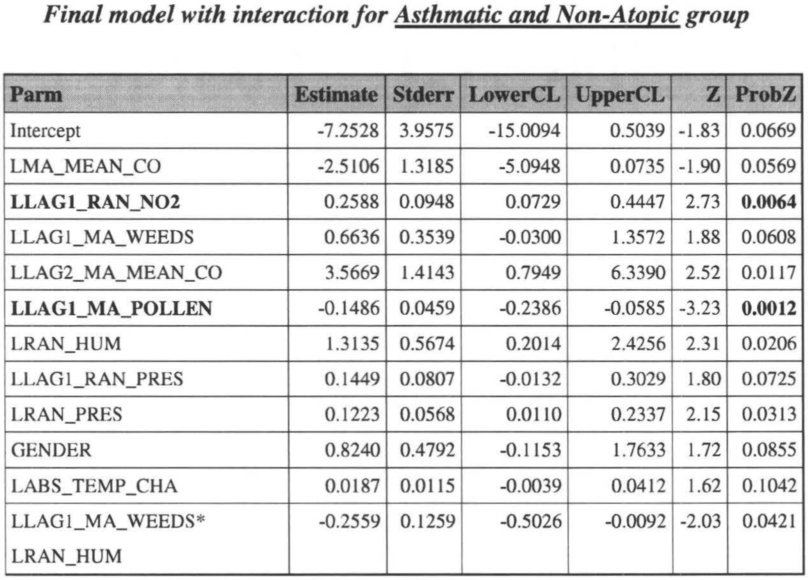
Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-36.1942	20.3033	-75.9880	3.5996	-1.78	0.0746
LMEAN_SO2	-0.0639	0.0308	-0.1242	-0.0036	-2.08	0.0378
LLAG2_MA_BASIDIO	-0.0225	0.0152	-0.0523	0.0072	-1.48	0.1377
LMA_ASCOM	-0.0548	0.0355	-0.1242	0.0147	-1.54	0.1225
LLAG2_RAN_PRES	0.1066	0.0598	-0.0106	0.2237	1.78	0.0746
LLAG2_MEAN_PRES	7.6139	4.4228	-1.0545	16.2823	1.72	0.0852

Figure 8.5: *Results for Asthmatic and Non-Atopic Group*



Final model without interaction for Asthmatic and Non-Atopic group

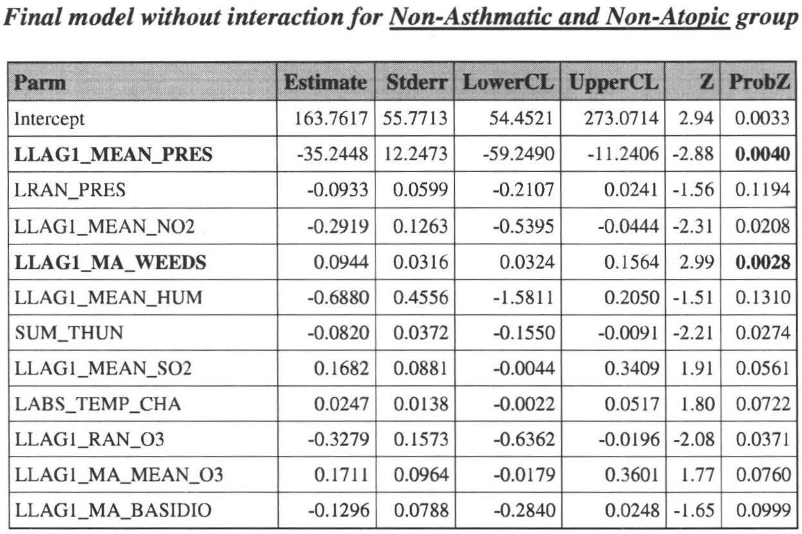
Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-4.0459	2.9049	-9.7394	1.6475	-1.39	0.1637
LMA_MEAN_CO	-2.4466	1.3263	-5.0461	0.1530	-1.84	0.0651
LLAG1_RAN_NO2	0.2640	0.0960	0.0759	0.4522	2.75	**0.0060**
LLAG1_MA_WEEDS	-0.1287	0.0760	-0.2777	0.0203	-1.69	0.0906
LLAG2_MA_MEAN_CO	3.6243	1.4159	0.8491	6.3995	2.56	0.0105
LLAG1_MA_POLLEN	-0.1537	0.0445	-0.2409	-0.0665	-3.46	**0.0005**
LRAN_HUM	0.2257	0.0941	0.0413	0.4101	2.40	0.0164
LLAG1_RAN_PRES	0.1620	0.0825	0.0003	0.3237	1.96	0.0496
LRAN_PRES	0.1167	0.0559	0.0072	0.2263	2.09	0.0368
GENDER	0.8460	0.4811	-0.0969	1.7889	1.76	0.0786
LABS_TEMP_CHA	0.0169	0.0108	-0.0043	0.0380	1.56	0.1181



Final model with interaction for Asthmatic and Non-Atopic group

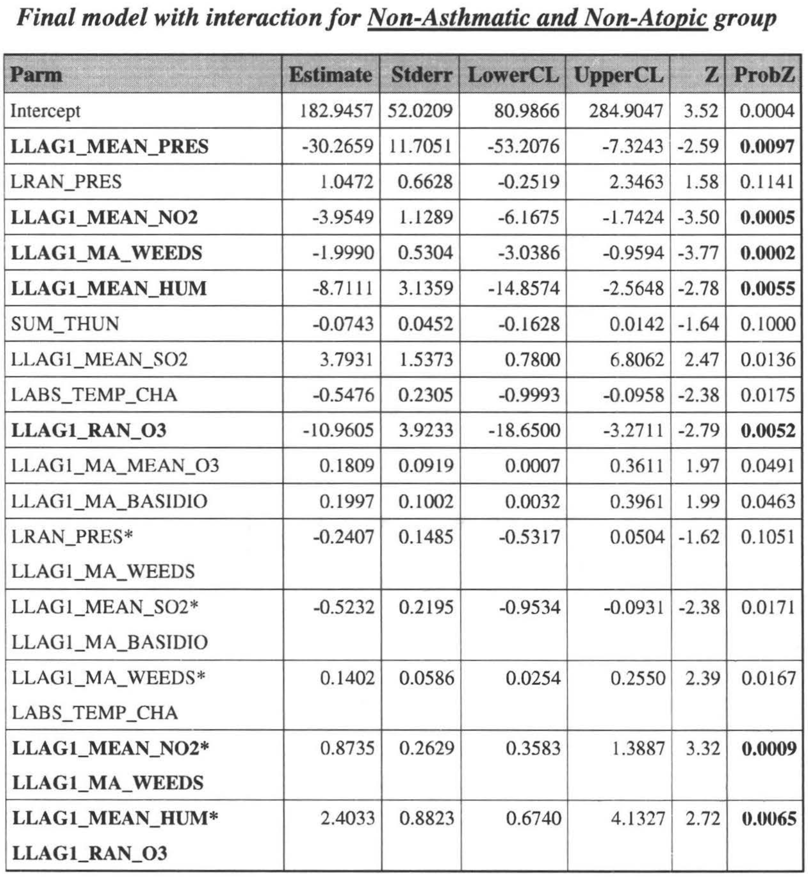
Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	-7.2528	3.9575	-15.0094	0.5039	-1.83	0.0669
LMA_MEAN_CO	-2.5106	1.3185	-5.0948	0.0735	-1.90	0.0569
LLAG1_RAN_NO2	0.2588	0.0948	0.0729	0.4447	2.73	**0.0064**
LLAG1_MA_WEEDS	0.6636	0.3539	-0.0300	1.3572	1.88	0.0608
LLAG2_MA_MEAN_CO	3.5669	1.4143	0.7949	6.3390	2.52	0.0117
LLAG1_MA_POLLEN	-0.1486	0.0459	-0.2386	-0.0585	-3.23	**0.0012**
LRAN_HUM	1.3135	0.5674	0.2014	2.4256	2.31	0.0206
LLAG1_RAN_PRES	0.1449	0.0807	-0.0132	0.3029	1.80	0.0725
LRAN_PRES	0.1223	0.0568	0.0110	0.2337	2.15	0.0313
GENDER	0.8240	0.4792	-0.1153	1.7633	1.72	0.0855
LABS_TEMP_CHA	0.0187	0.0115	-0.0039	0.0412	1.62	0.1042
LLAG1_MA_WEEDS* LRAN_HUM	-0.2559	0.1259	-0.5026	-0.0092	-2.03	0.0421

Figure 8.6: *Results for Non-Asthmatic and Non-Atopic Group*



Final model without interaction for <u>Non-Asthmatic and Non-Atopic</u> group

Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	163.7617	55.7713	54.4521	273.0714	2.94	0.0033
LLAG1_MEAN_PRES	-35.2448	12.2473	-59.2490	-11.2406	-2.88	**0.0040**
LRAN_PRES	-0.0933	0.0599	-0.2107	0.0241	-1.56	0.1194
LLAG1_MEAN_NO2	-0.2919	0.1263	-0.5395	-0.0444	-2.31	0.0208
LLAG1_MA_WEEDS	0.0944	0.0316	0.0324	0.1564	2.99	**0.0028**
LLAG1_MEAN_HUM	-0.6880	0.4556	-1.5811	0.2050	-1.51	0.1310
SUM_THUN	-0.0820	0.0372	-0.1550	-0.0091	-2.21	0.0274
LLAG1_MEAN_SO2	0.1682	0.0881	-0.0044	0.3409	1.91	0.0561
LABS_TEMP_CHA	0.0247	0.0138	-0.0022	0.0517	1.80	0.0722
LLAG1_RAN_O3	-0.3279	0.1573	-0.6362	-0.0196	-2.08	0.0371
LLAG1_MA_MEAN_O3	0.1711	0.0964	-0.0179	0.3601	1.77	0.0760
LLAG1_MA_BASIDIO	-0.1296	0.0788	-0.2840	0.0248	-1.65	0.0999



Final model with interaction for <u>Non-Asthmatic and Non-Atopic</u> group

Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ
Intercept	182.9457	52.0209	80.9866	284.9047	3.52	0.0004
LLAG1_MEAN_PRES	-30.2659	11.7051	-53.2076	-7.3243	-2.59	**0.0097**
LRAN_PRES	1.0472	0.6628	-0.2519	2.3463	1.58	0.1141
LLAG1_MEAN_NO2	-3.9549	1.1289	-6.1675	-1.7424	-3.50	**0.0005**
LLAG1_MA_WEEDS	-1.9990	0.5304	-3.0386	-0.9594	-3.77	**0.0002**
LLAG1_MEAN_HUM	-8.7111	3.1359	-14.8574	-2.5648	-2.78	**0.0055**
SUM_THUN	-0.0743	0.0452	-0.1628	0.0142	-1.64	0.1000
LLAG1_MEAN_SO2	3.7931	1.5373	0.7800	6.8062	2.47	0.0136
LABS_TEMP_CHA	-0.5476	0.2305	-0.9993	-0.0958	-2.38	0.0175
LLAG1_RAN_O3	-10.9605	3.9233	-18.6500	-3.2711	-2.79	**0.0052**
LLAG1_MA_MEAN_O3	0.1809	0.0919	0.0007	0.3611	1.97	0.0491
LLAG1_MA_BASIDIO	0.1997	0.1002	0.0032	0.3961	1.99	0.0463
LRAN_PRES* LLAG1_MA_WEEDS	-0.2407	0.1485	-0.5317	0.0504	-1.62	0.1051
LLAG1_MEAN_SO2* LLAG1_MA_BASIDIO	-0.5232	0.2195	-0.9534	-0.0931	-2.38	0.0171
LLAG1_MA_WEEDS* LABS_TEMP_CHA	0.1402	0.0586	0.0254	0.2550	2.39	0.0167
LLAG1_MEAN_NO2* LLAG1_MA_WEEDS	0.8735	0.2629	0.3583	1.3887	3.32	**0.0009**
LLAG1_MEAN_HUM* LLAG1_RAN_O3	2.4033	0.8823	0.6740	4.1327	2.72	**0.0065**

Bibliography

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

[2] Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd edition. NewYork: Wiley.

[3] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

[4] Wei Pan (2001). Akaike's Information Criterion in Generalized Estimating Equation. *Biometrics 2001*, 57: 120-125

[5] Barnhart, H. X. and Williamson, J. M. (1998). Goodness-offit tests for GEE modeling with binary data. *Biometrics*, 54:720-729

[6] Gary A. Ballinger(2004). Using Generalized Estimaitng Equation for longitudinal data analysis. *Organizational Research Methods 2004*, Vol. 7 No.2: 127-150

[7] Christopher J. W. Zorn. (2001). Generalized Estimating Equations Models for correlated data: A review with Applications. *American Journal of Political Science 2004*, Vol. 45, No 2: 470-490.

[8] Davidian, M. and Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data.* London: Chapman and Hall/CRC Press.