STATISTICAL ANALYSIS OF ASTHMA HOSPITALIZATION INCIDENCES IN CANADIAN CHILDREN

STATISTICAL ANALYSIS OF ASTHMA HOSPITALIZATION INCIDENCES IN CANADIAN CHILDREN

By

JENNIFER DAI, B.Sc.

A Project

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

December 2002

MASTER OF SCIENCE (2002) **McMaster University**
(Statistics) **Hamilton, Ontario**

TITLE: **Statistical Analysis of Asthma Hospitalization Incidences in Canadian Children**

AUTHOR: **Jennifer Dai, B.Sc.**

SUPERVISOR: **Dr. Román Viveros-Aguilera**

NUMBER OF PAGES: **xi, 60**

Abstract

Asthma is the leading chronic disease of children in industrialized countries. In Canada, it is the most common cause of hospital admissions in children. Data were assembled for all asthma hospitalizations in Canada from 1990 to 2000 by the Canadian Institute for Health Information (CIHI). The annual cycles of asthma hospitalization among Canadian children from 1990 to 2000 were compared. For every year, region and latitude, asthma hospitalizations were lowest in July and August followed by a major peak in September then a rapid decline.

Contingency table analyses were done to examine the homogeneity of the distributions of asthma hospitalization counts for the factors age, gender, region and latitude groups. Age, region and latitude groups were found to be significantly different with respect to their distribution of asthma counts. However, the distributions of asthma hospitalization counts did not differ significantly for gender.

A nonlinear least squares model was fitted to the asthma hospitalization data for weeks 30 to 42. The primary objective was to obtain estimates of the parameter that describes the timing of the September peak. Next, a likelihood ratio test was done to assess the homogeneity of the September peaks for the factors age, gender, latitude and region. We found that, apart from gender, the

September peaks were significantly different. Furthermore, the annual cycle of asthma hospitalization for children aged 2 to 4 was identical to that of children aged 5 to 15 except the peak in hospitalization for 2 to 4 year olds occurred on average 2 days after the older children. We suspect that the increase prevalence of and exposure to viral infections, exposure of school-aged children to allergens at school and the transmission of these factors to younger siblings are responsible for the September asthma epidemic.

A Quasi-Poisson log-linear model was also fitted to the data to assess jointly the effects of age, gender, latitude, year and risk group size. The data were overdispersed, after accounting for overdispersion, we found that age, gender, latitude, year and the interactions between age and gender, age and latitude and gender and latitude were significant in explaining the data. Surprisingly, time had a negative effect suggesting a tendency to decline in the number of asthma incidences requiring hospitalization over the years.

Acknowledgements

This project could have never been done without the guidance, help and expertise of many people. I am grateful to have this opportunity to thank them in writing. Dr. Román Viveros-Aguilera has been a fantastic supervisor throughout the entire process. I was very fortunate to have his understanding, patience and excellent guidance. Many thanks to Dr. Peter Macdonald for his expertise and his fresh perspective when it was needed. Thanks also to Mr. Neil Johnston for his insightful discussions and continual support both financially and through technical advice. I am grateful to Dr. Geoff Norman for his help during the initial stages of my project. A special thank you to Dr. Aaron Childs for introducing me to Mr. Neil Johnston, for listening patiently to my many concerns and for his valuable recommendations.

In addition, I would like to extend my appreciation for the insightful comments from my defence committee which consisted of Dr. Aaron Childs, Dr. Lehana Thabane and Dr. Román Viveros-Aguilera as well as the support from the Department of Mathematics and Statistics.

I would like to thank the extraordinary group of people I had the good fortune to have as colleagues throughout my career as a graduate student. They have been the most wonderful study partners and support group. Their friendship alone made this experience worthwhile for me.

Very special thanks go to Mom, Dad, Carolanne, Peter and Hristo for their unwavering faith and encouragement throughout the many twists and turns in my life.

TABLE OF CONTENTS

Page

CHAPTER 1: The National Asthma Hospitalization Data and Study Objectives

Chapter 2: Contingency Table Analyses of the Asthma Hospitalization Data

Chapter 3: Non-Linear Least Squares Curve Fitting for September Peaks

Chapter 4: Likelihood Ratio Test for Homogeneity of the September Peak Locations

Chapter 5: General Linear Modelling of Pediatric Asthma Hospitalization Data

Chapter 6: Conclusions and Discussion

Appendix

List of Figures

List of Tables

Chapter 1: The National Asthma Hospitalization Data and Study Objectives

1.1 The National Asthma Hospitalization Data

Asthma is a common respiratory disorder characterized by difficulty in breathing, wheezing and a sense of constriction in the chest. The World Health Organization considers asthma a serious public health problem with over 100 million sufferers worldwide. Asthma is the leading chronic disease of children in industrialized countries and the most common childhood illness after the common cold. It is also the leading cause of absenteeism from school. In Canada, 10 to 15 percent of children are reported to have asthma. It is the number one cause of hospitalization for Canadian children (Health Canada 2001).

Twenty percent of all annual asthma hospitalizations of children between ages 5 and 15 occur in September (Johnston and Sears 2001). This trend has been reported in many Northern Hemisphere countries such as Canada, the United States, the United Kingdom, Ireland, Mexico, Hong Kong, Greece, Israel, Finland and Trinidad. There is no indication in the Southern Hemisphere of the same phenomenon. The September peak in asthma hospitalization for children aged 5 to 15 may be attributed to changes in weather patterns, ambient air pollution levels, start-up of heating systems, viral infections and seasonal allergen levels. However, no conclusive aetiology has been established.

Previous reports of the September peak have been based on case series or data from small geographic areas with little variation in climate or environmental factors. Environmental variation may exacerbate the risk of asthma. Hospitalization data are available for the population of Canada, distributed over more than 85 degrees of longitude and 30 degrees of latitude. Hence the quality and wide coverage of data has created an opportunity to study the timing of the September increase in areas with a variety of environments.

Data were assembled for all asthma hospitalizations in Canada from 1990 to 2001. The Canadian Institute for Health Information (CIHI) collects data from hospitals across Canada in a standardized form for the Provincial Governments. All records of hospital discharges containing a responsible diagnosis of asthma (ICD-9 493) from April 1990 to March 2001 were selected. The data records provided a consistently scrambled identity number that allowed multiple events for the same patient to be tracked. The data included individual postal or similar codes. Standard conversion files (Mapinfo Canada Inc.) were used to assign the data to regions or latitudes of residence. Table 1.1 describes in detail the seven variables made available to us from the National Asthma Hospitalization Data. In their raw format, the data formed a 326 991 by 7 matrix with each row corresponding to a hospitalization count of asthma attacks. The columns correspond to the variables available to us from the study.

The data for the current study were obtained from Mr. Neil Johnston, a research member of the Firestone Institute for Respiratory Illness at St. Joseph's Hospital in

Hamilton, Ontario. Mr. Johnston purchased and obtained permission to use the National Asthma Hospitalization Data for the purpose of studying pediatric asthma in Canada. He enlisted the help of Dr. Román Viveros, Dr. Geoff Norman and myself to pursue his research objectives. Some discussions with Dr. Norman and many more with Mr. Johnston helped to clarify the aims of the study and to guide some of the statistical modelling and analyses undertaken.

The data came in an Excel file. SPlus and SPSS versions were made for the statistical analyses. About 3% of the subjects had missing values, primarily on the latitude and longitude entries.

Table 1.1 The seven variables available from the National Asthma Hospitalization Data.

Variable Name	**Description**	**Range**
Age	The age of patients hospitalized for asthma.	[2, 109]
Gender	The gender of patients hospitalized for asthma.	Male, Female
Year	The year when the asthma hospitalization occurred.	1990-2001
Month	The month of the year when the asthma hospitalization occurred.	[1, 12]
Day	The day of the month when the asthma hospitalization occurred.	[1, 31]
Latitude	The latitude of the place of resident where the asthma sufferer lived.	[41.76, 73.02]
Longitude	The longitude of the place of resident where the asthma sufferer lived.	[-52.67, -140.88]

Dates of hospital admission were assigned to a week of the year for analysis to eliminate the effects of variation in hospitalization patterns by day of the week. This was done by counting consecutive seven day periods from January 1st in each year not including February 29th. The sums of admissions for the eight days from December 24th to December 31st were reduced by 12.5% to account for the odd day.

Children aged 2 through 15 between April 1, 1990 and March 31, 2001 were chosen for the study. The potential differences in the annual hospitalization cycle for children having one, two to four or over four admissions during the study period were examined. The timing of the September peak was identical at all three frequency groups. As a result, all admissions for pediatric asthma were included in the analyses as single events, whether they be a first or repeat admission.

1.2 Case Selection and Variable Grouping

After removing the cases with missing values, we retained for the study 86 435 cases in the 2-15 age group. In consultation with Mr. Johnston, some groupings were undertaken. Table 1.2 describes the variable groupings considered in the study. Several cross groupings were also examined.

Table 1.2 Variable Groupings in Study.

Variable	Number of Groups	Groups
Age	2	Ages 2-4 and 5-15
Gender	2	Males and Females
Region	4	British Columbia, Prairie, Ontario and Atlantic Regions
Latitude	4	Under 44, 44-50, 51-52 and Over 52

1. Age

Asthma diagnoses for infants less than 24 months of age are highly variable and often difficult to distinguish from respiratory tract infections. Hence they were excluded from further analyses. We were interested in examining children aged 2 to 4 and 5 to 15.

2. Gender

The natural grouping Male/Female was considered.

3. Region

Canada was divided into four regions for this study. They were Atlantic Provinces (Newfoundland, Prince Edward Island, Nova Scotia and New Brunswick), Ontario, the Prairies (Manitoba, Saskatchewan and Alberta) and British Columbia. The data for Quebec were not directly collected by the Canadian Institute for Health Information. The data available for Quebec contained assigned age groups not consistent with those of interest in the study. The data for the province of Quebec revealed similar asthma cycles to the data grouped in a similar manner for the rest of Canada. Hence Quebec was excluded from the analyses.

4. Latitude

Ranges of Latitude were chosen to divide the population of Canada living close to the United States border into two approximately equal portions and also to divide the population living above latitude 51 into two smaller approximately equal sized populations. This resulted in four latitude groups altogether. The four groups were latitudes under 44, latitude between 44 and 50, latitudes between 51 and 52 and latitudes over 52.

1.3 Objectives of the Study in this Project

The overall objective is to characterize the temporal, spatial as well as age and gender patterns of pediatric asthma hospitalization in Canada. The more specific objectives are:

a) To assess individually the effects of age, gender, latitude, region and year on the incidences of asthma attacks.

b) To develop a statistical model for the September peaks and use the model to estimate the peaks' locations.

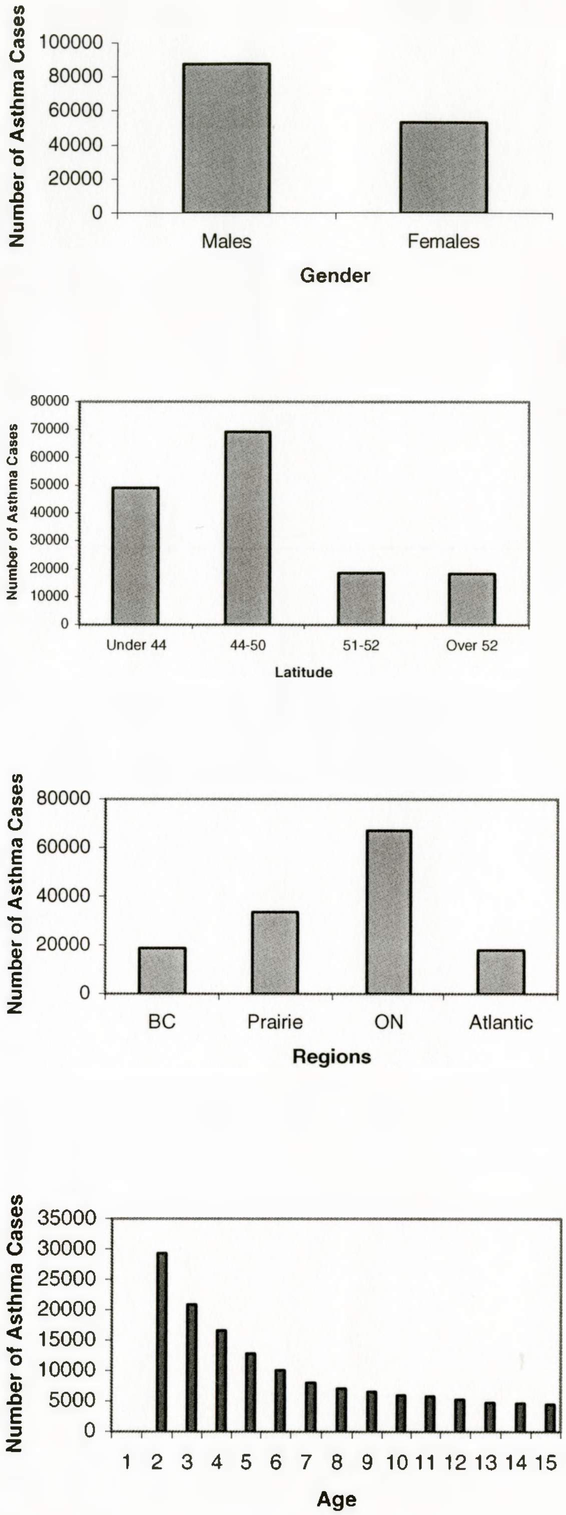
c) To assess the consistency (agreement) of the peaks' locations across age, gender, latitudes and regions.

d) To develop a regression model to estimate and assess the joint effects of the factors considered aimed at discarding irrelevant factors in the presence of the others.

1.4 Descriptive Analyses of the National Asthma Hospitalization Data

Summary bar plots for the factors considered are presented in Figure 1.1. Note these plots purely reflect the raw information in the data, no population adjustments has been done. There were 24.4% more males aged 2 to 15 in the study than females. In terms of latitude groupings, most of the asthma cases fell in the 44 to 50 latitude group. Furthermore, Ontario had the most asthma hospitalizations for children aged 2 to 15. Children aged 2 to 4 made up 46.8% of the study. Note a fairly regular curved decay trend in the number of asthma attacks as the patients get older.

Figure 1.1 Bar Plots of Asthma Hospitalization Counts.



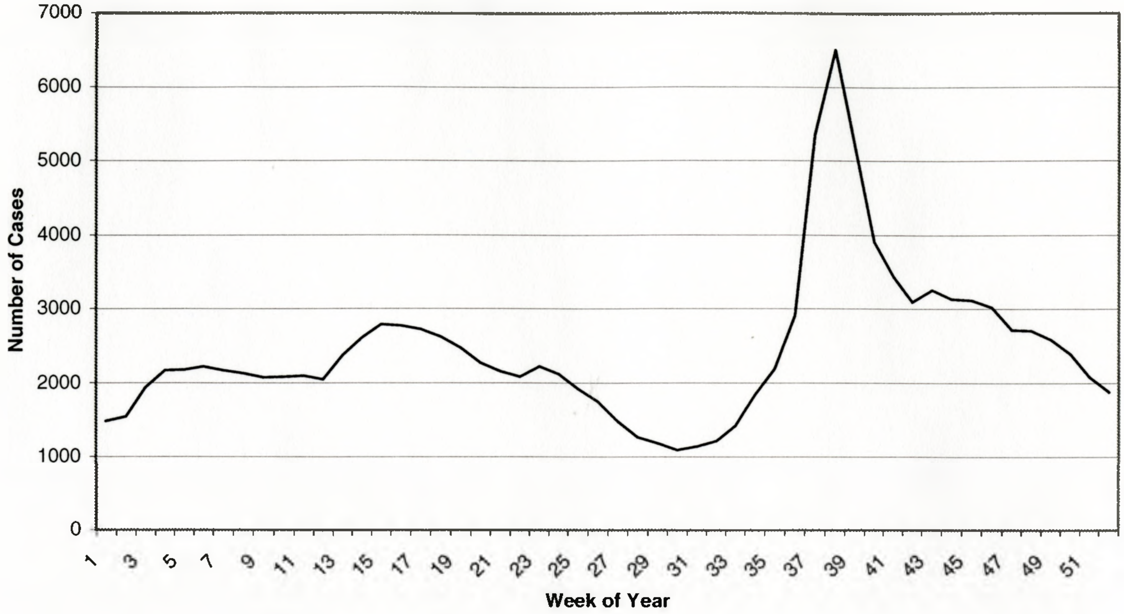


Figure 1.2. Plot of WeeklyAsthma Hospitalization Counts of Children Aged 2 to 15 from 1991 to 2000.

Figure 1.2 depicts the cycle of asthma hospitalization of Canadian children aged 2 to 15 from 1991 to 2000. The ten-year period is aggregated into a single annual cycle. The cycle shows an initial slow rise in the number of admissions from winter to spring. This is followed by a decline through the month of June. A low point is reached in week 30 (July 23 to 30). The weekly number of cases then rises to reach double the number of admissions compared to the low point in week 30 by week 35 (August 27 to September 2) before Labour Day. Next, the curve accelerates to a maximum peak at week 38 (September 17 to 23) followed by a rapid decline to week 42 (October 15 to 21). From this point, the curve slowly declines to the end of the year. Figure 1.3 shows the number of pediatric asthma admissions per week for each year from 1990 to 2001. The annual cycles show a similar pattern to the one described in Figure 1.2 for each year. The September peak occurs in week 38 for every year except in 1992 when it occurred in week 39 and in 1997 when the peak was at week 37. The trough in the annual cycles consistently occurred between weeks 29 and 33. In addition, the doubling of the lowest value consistently occurred immediately before Labour Day.

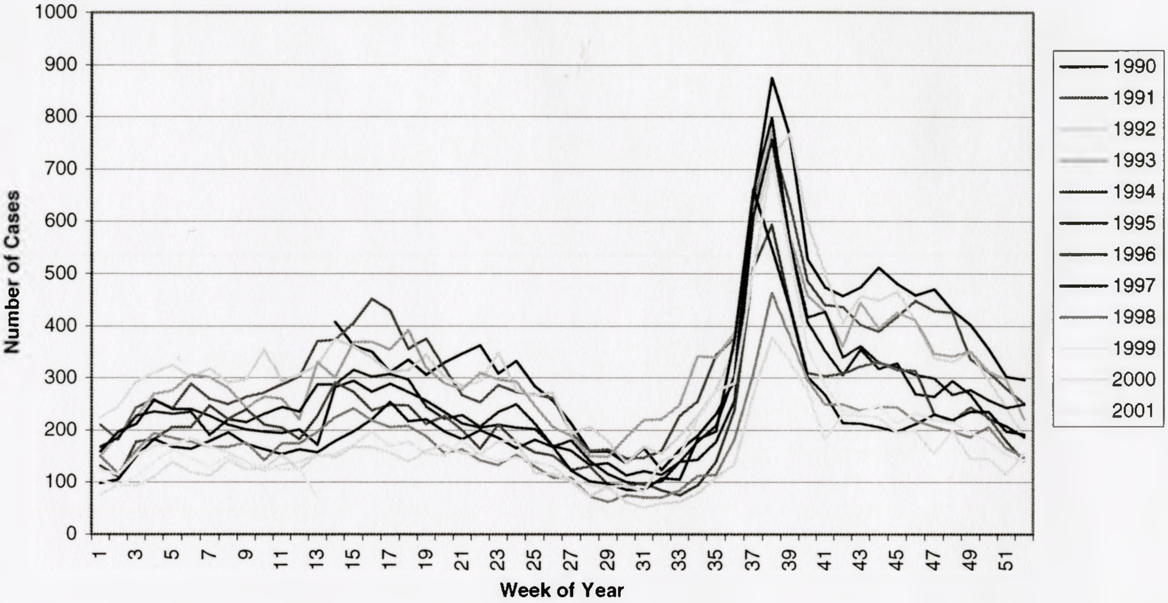


Figure 1.3. Plot of Asthma Hospitalization Counts by Year for Children Aged 2 to 15.

Figure 1.4 shows the annual asthma hospitalization cycle of pre-school (2 to 4 years old) and school aged (5 to 15 years old) children aggregated over a ten year period from 1991 to 2000. The cycles appear to be consistent with that in Figure 1.2 in both pre-school and school aged children with the exception that the overall low point in the pre-school group is week 31 rather than week 30. Figures 1.5 and 1.6 show the asthma admission cycles from 1990 to 2001 for pre-school and school aged children respectively. Comparing Figure 1.4 with Figure 1.6 reveals that the timing of the cycles for both age groups is essentially the same across the eleven years. However, the intensity of the September peak varies. The intensity of the September peak is greater in the school aged children.

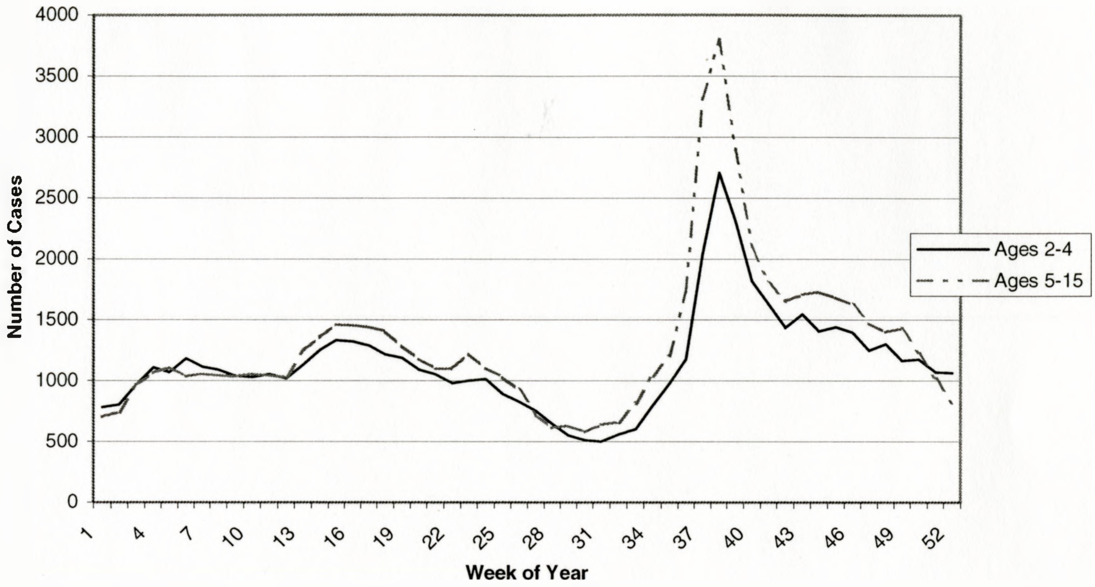


Figure 1.4. Plot of Weekly Asthma Hospitalization Counts in Children Aged 2 to 4 and 5 to 15 from 1991 to 2000.

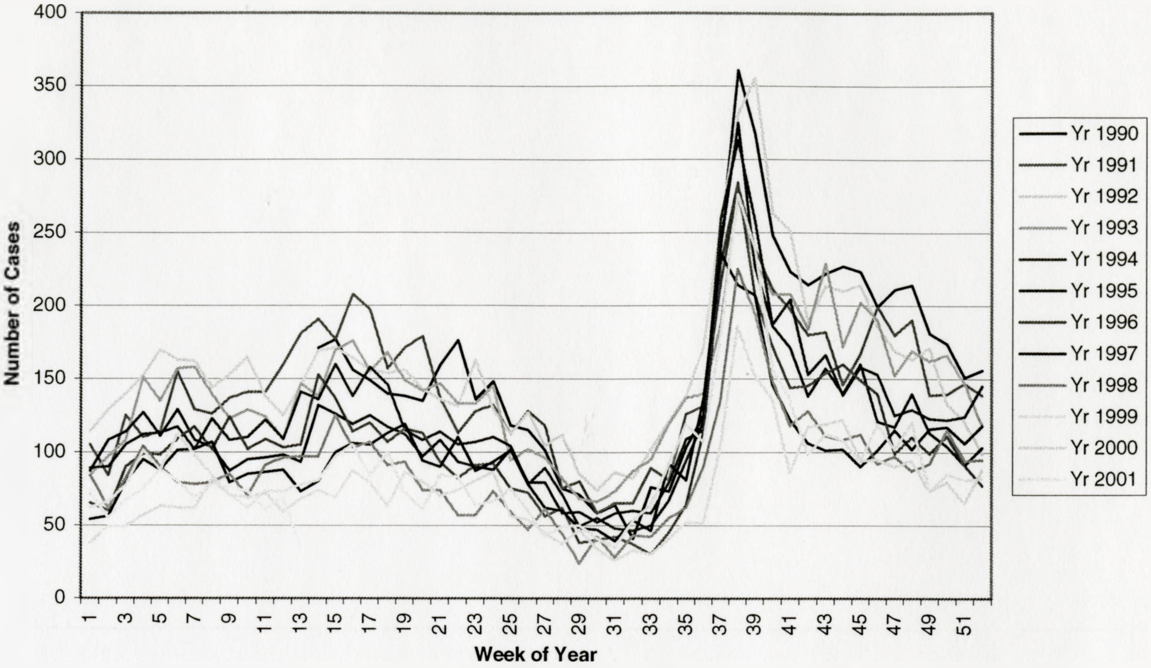


Figure 1.5. Plot of the Weekly Asthma Hospital Admissions for Children Aged 2 to 4 from 1990 to 2001.

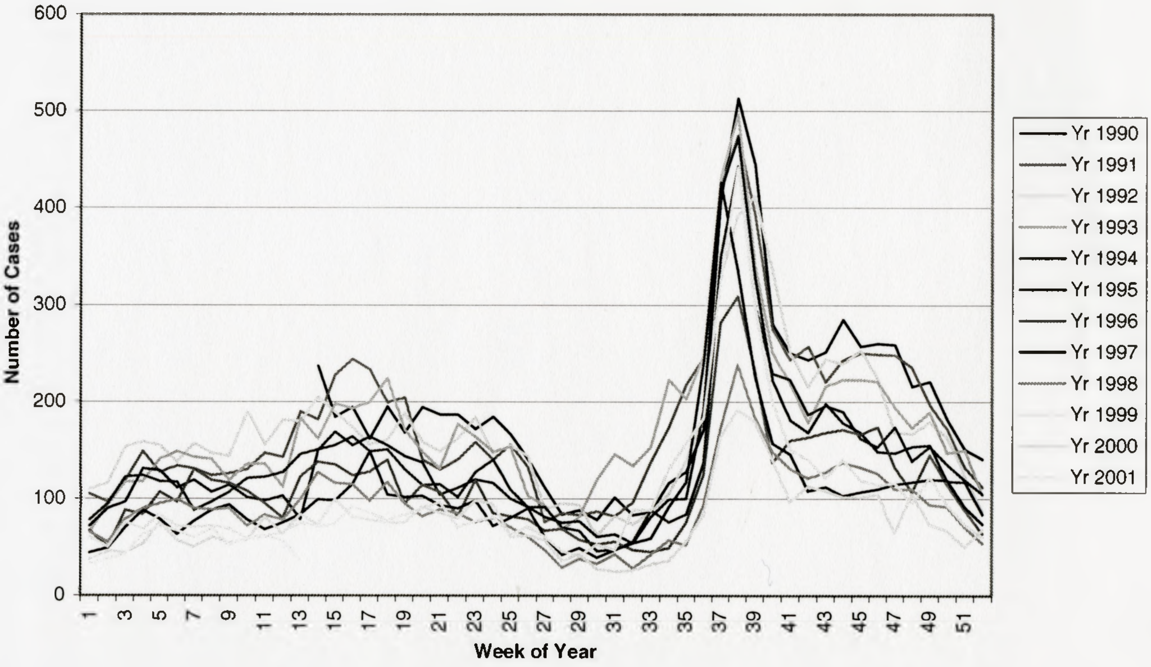


Figure 1.6. Plot of the Weekly Asthma Hospital Admissions for Children Aged 5 to 15 from 1990 to 2001.

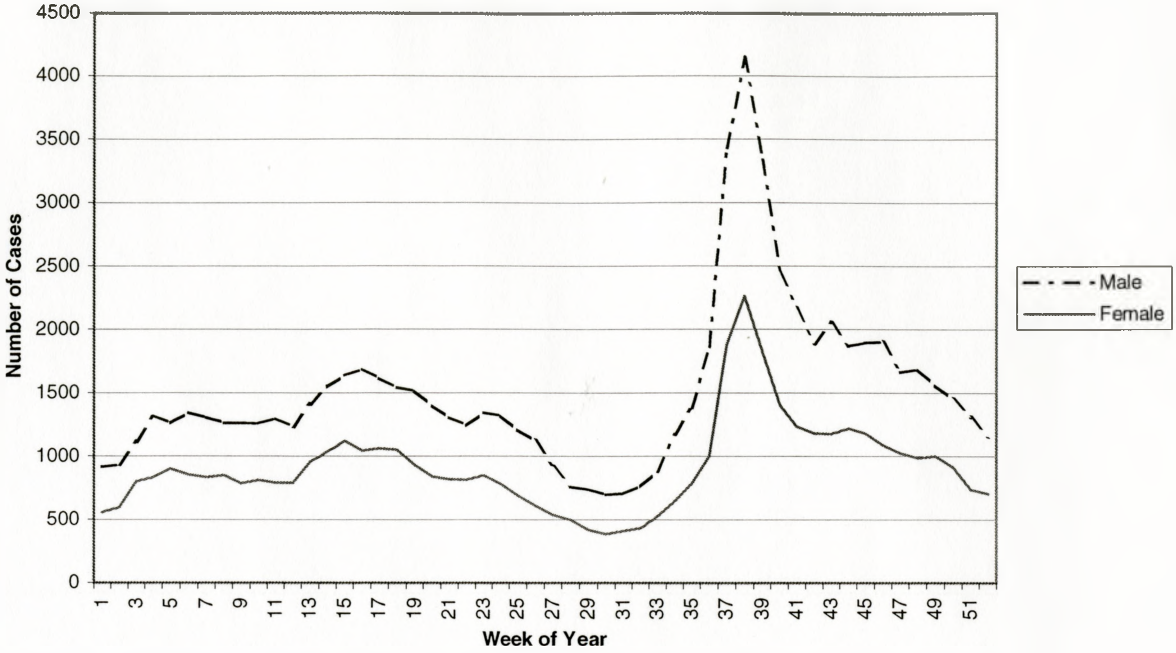


Figure 1.7. Plot of the Weekly Asthma Hospital Admissions for Males and Females Aged 2 to 15 from 1991 to 2000.

Figure 1.7 shows that while males aged 2 to 15 have approximately double the risk of hospitalization of females, the annual asthma cycles for both genders are identical to the one described in Figure 1.2.

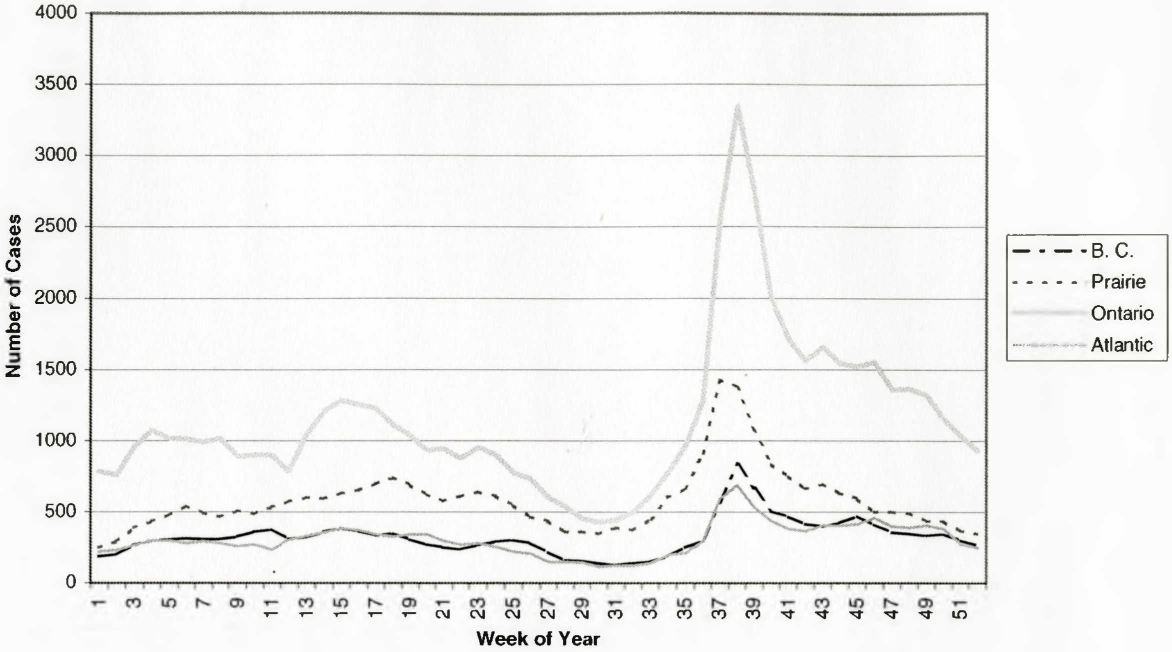


Figure 1.8. Plot of the Weekly Asthma Admissions for Children Aged 2 to 15 from 1991 to 2000 Across the Four Regions of Canada.

Figure 1.8 shows the aggregated ten year cycle of asthma hospitalization for children aged 2 to 15 in the four regions of Canada from 1991 to 2000. The climate varies considerably between British Columbia, the Prairies, Ontario and the Atlantic Provinces. The timing of the peak and trough for the four regions is consistent with the annual pediatric asthma hospitalization cycle described earlier in Figure 1.2. However, the peak intensity varies considerably between the regions as well as other trends during different times of the year.

Figure 1.9 shows the aggregated ten year cycle of asthma hospitalization for children aged 2 to 15 by range of latitude of residence from 1991 to 2000. There is consistency in the timing of the September peak and variations in the peak intensity in all four latitude groupings. Also, the cycle of asthma hospitalization of the four latitudes is consistent with the one described early in Figure 1.2. Note the closeness of the curves for the 51-52 and above 52 latitude groups.

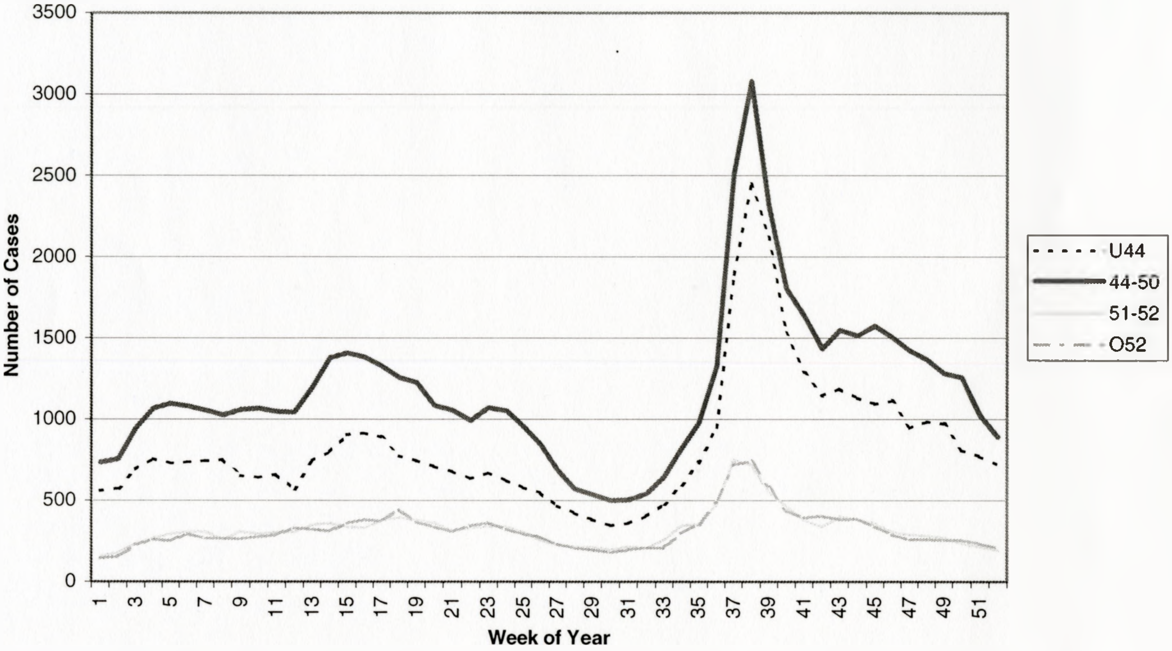


Figure 1.9. Plot of the Weekly Asthma Admissions for Children Aged 2 to 15 from 1991 to 2000 Over the Four Latitude Groupings.

Chapter 2: Contingency Table Analyses of the Asthma Hospitalization Data

2.1 Contingency Table Analysis

A contingency table is a way of summarising categorical data and describing the relationship between variables. It is a table of frequencies obtained by classification of individuals or units according to the values of the variables in question.

One important use of contingency tables occurs when we have a number of populations and we wish to assess whether there is agreement in the probabilities of classification of a given variable across the populations. This method was applied in our project. This application is often referred to as a test for *homogeneity* across the populations.

Consider the case of c populations labeled as $1,2,\ldots,c$. Suppose the variable considered has r levels, labeled as $1,2,\ldots,r$. Denote by P_{ij} the probability that for population *j*, the variable takes on the i^{th} level. This probability is sometimes denoted by $P_{i/j}$. For n individuals randomly selected, with n_j of them coming from the j^{th} population, the observed frequencies $(n_{1j}, n_{2j}, \ldots, n_{rj})$ have a multinomial distribution with index n_j and probability parameters $(P_{1j}, P_{2j}, \ldots, P_{rj})$. Note that $\sum_{i=1}^{r} P_{ij} = 1$. In the context of testing statistical hypotheses, our problem is to test the null hypothesis of homogeneity, namely

$$H_o : P_{i1} = P_{i2} = \cdots = P_{ic},\ 1 \le i \le r.$$

Conditioning on the sample sizes $n_1, n_2, \ldots, n_c$, the maximum likelihood estimates under the assumption of homogeneity are

$$\hat{P}_{ij} = \frac{n_{i+}}{n},\quad j = 1,2,\ldots,c\ ,\quad i = 1,2,\ldots,r$$

where $n_{i+} = \sum_{j=1}^{c} n_{ij}$. Thus the estimated expected frequency under homogeneity for the (i, j) cell is

$$\hat{e}_{ij} = n_j \hat{P}_{ij} = \frac{n_{i+}n_{+j}}{n}\ ,\quad 1 \le j \le c,\ \ 1 \le i \le r$$

where $n_j = n_{+j} = \sum_{i=1}^{r} n_{ij}$.

The simplest and best known test for the assumption of homogeneity is Pearson's Chi-squared test based on the statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}.$$

χ^2 is a measure of agreement between the observed and estimated expected frequencies with large values indicating disagreement with H_o. If the assumption of homogeneity H_o is true then χ^2 has an approximate Chi-squared distribution with $(r-1)(c-1)$ degrees of freedom (df). The approximation improves as n gets larger with the rule of thumb being that each $\hat{e}_{ij}$ should be at least 5 for an acceptable approximation. The likelihood ratio test is an alternative, also having an approximate Chi-squared distribution. For details, see Lindgren (1993, Chap. 10).

2.2 Application of Contingency Table Analysis to the Asthma Hospitalization Data

Two-way contingency tables were constructed for counts of asthma cases for each of the factors age, gender, region and latitude over the 52 weeks in a year. Thus the factor levels define the populations and the weeks of the years are levels of the variable for the classification within each population. The data from years 1991, 1995 and 2000 were chosen for each table. The levels of the factors correspond to the groupings in Table 1.2 (Chapter 1). The two-way contingency tables for the factors age and gender consisted of 52 rows and 2 columns. For the factors region and latitude, the two-way contingency tables had 52 rows and 4 columns. Test for homogeneity was carried out for each of the factors in order to assess the homogeneity of the weekly distribution across each factor's grouping. In other words, we want to determine whether the weekly proportions are the same across the levels of each factor. The chi-square test statistic was calculated for each table and a p-value was obtained. The S-Plus function given in Appendix A was written and used to carry out the calculations.

2.3 Results of the Contingency Table Analyses

Table 2.1 presents the results of the two-way contingency table analysis for each of the factors.

Table 2.1 Contingency Table Analysis Results.

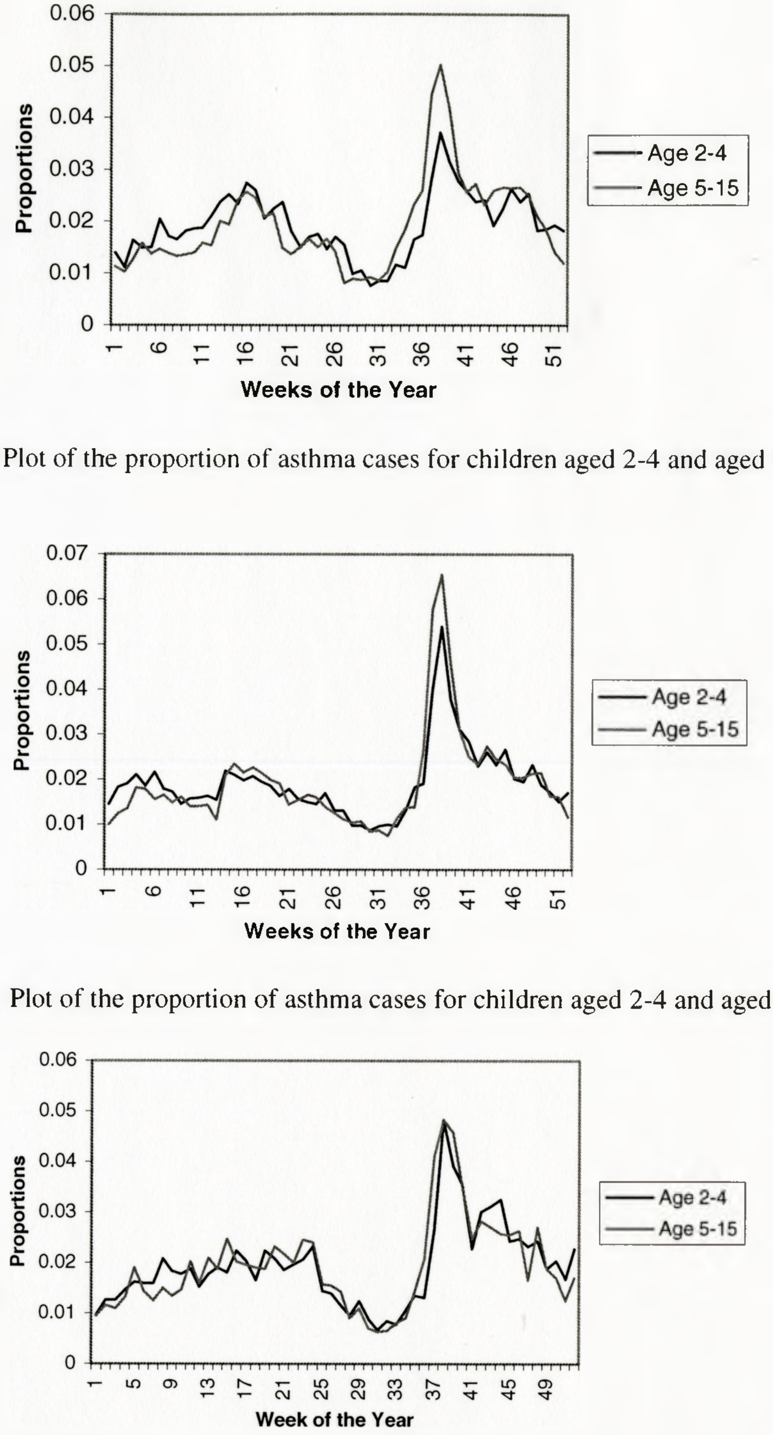
Year	1991			1995			2000		
	χ^2	(df)	P-value	χ^2	(df)	P-value	χ^2	(df)	P-value
Age	241.397	(51)	0.0000	110.580	(51)	0.0000027	65.143	(51)	0.0879
Gender	63.919	(51)	0.106	57.695	(51)	0.242	61.545	(51)	0.148
Region	820.913	(153)	0.0000	619.913	(153)	0.0000	275.406	(153)	0.0000
Latitude	586.547	(153)	0.0000	451.246	(153)	0.0000	239.308	(153)	0.00001

At a 5% level of significance, the p-values indicate that there is significant evidence of a difference between the distribution of asthma cases over the year for children aged 2 to 4 and children aged 5 to 15 in 1991 and 1995, but the evidence is borderline for 2000. The distributions of asthma cases for males and females over the year do not differ significantly in any of the three years considered. Significant differences in the weekly distribution of asthma cases were found across the four regions and across the four latitude groups for each of the three years considered.

2.4 Identifying Differences

A natural question is to find out where are the differences. By examining several appropriate plots, the areas for the lack of homogeneity may be identified. Figure 2.1 shows the observed proportion of asthma cases in preschool and school-aged children in the years 1991, 1995 and 2000. For the year 1991, the plot shows a clear tendency for smaller proportions in school-aged children up to week 34 compared to preschoolers while the tendency reverses from weeks 35 to 50. For year 1995, the main areas of differences occurred between weeks 1 and 21 and between weeks 36 and 41. The lack of homogeneity in year 2000 seemed to stem from the differences in the proportions from weeks 1 through 25 and from weeks 42 to 52. Figure 2.2 shows the proportion of asthma cases in children aged 2 to 15 grouped by latitude. The lack of homogeneity occurred at the peaks and between weeks 41 and 52 in 1991. The proportions for the four latitude groups appear to differ throughout the 52 weeks in 1995 and 2000.

Figure 2.1 Proportion of Asthma Cases in Children Aged 2 to 4 and 5 to 15 over the three years.

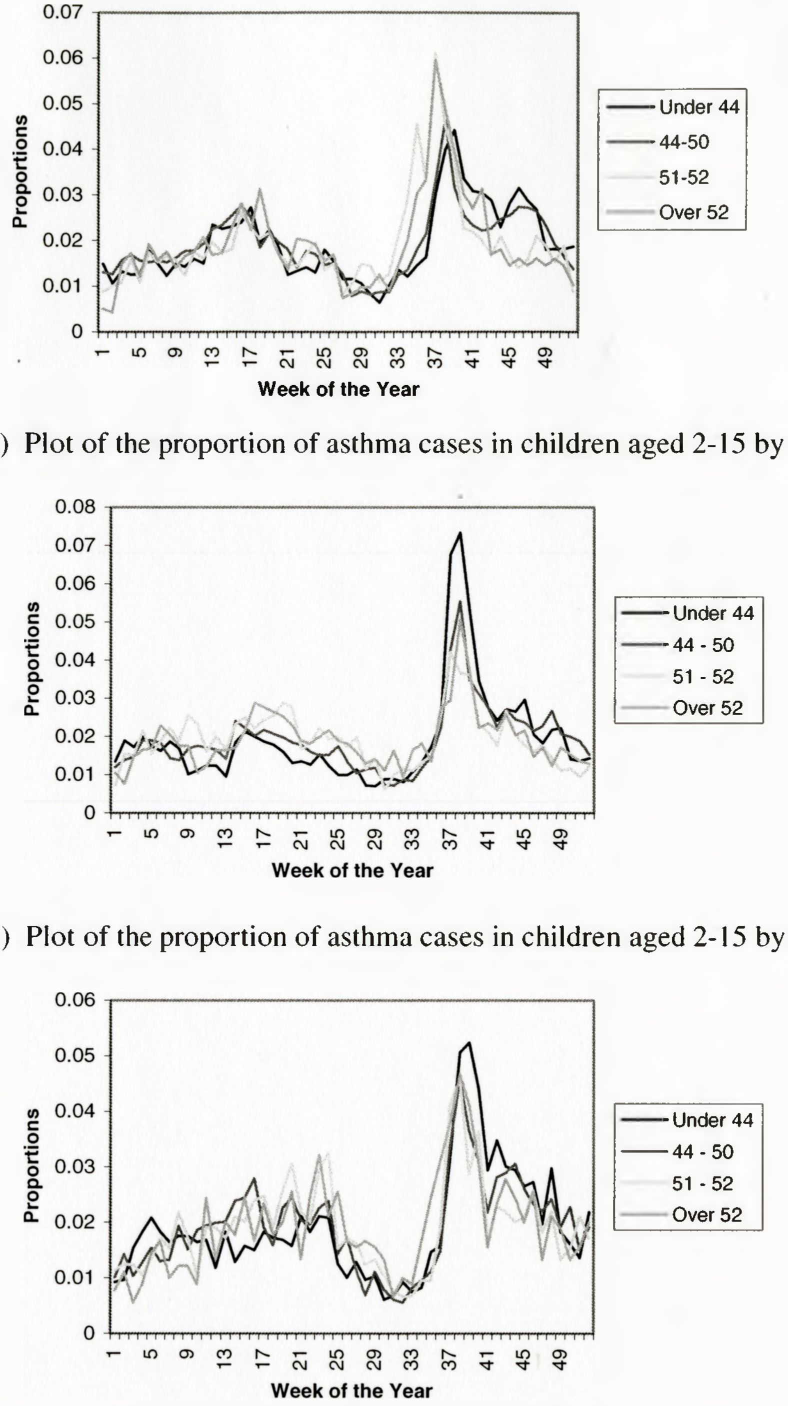


a) Plot of the proportion of asthma cases for children aged 2-4 and aged 5-15 in year 1991

b) Plot of the proportion of asthma cases for children aged 2-4 and aged 5-15 in year 1995

c) Plot of the proportion of asthma cases for children aged 2-4 and aged 5-15 in year 2000

Figure 2.2 Proportion of Asthma Cases by Latitude Groups Over Three Years.



a) Plot of the proportion of asthma cases in children aged 2-15 by latitude in 1991

b) Plot of the proportion of asthma cases in children aged 2-15 by latitude in 1995

c) Plot of the proportion of asthma cases in children aged 2-15 by latitude in 2000

Chapter 3: Nonlinear Least Squares Curve Fitting for September Peaks

3.1 Introduction to Nonlinear Least Squares Curve Fitting

Given a set of observations, it is often desirable to condense and summarize the data by fitting a model that depends on adjustable parameters. The basic approach to modelling outlined in Lawson and Hanson (1974) involves selecting or designing a merit function that measures the agreement between the data and the model with a particular choice of parameters. The merit function is arranged so that small values represent close agreement. The parameters of the model are then adjusted to achieve a minimum in the merit function, yielding the best-fit parameters. The adjustment process usually involves minimization in multi-dimensions. Ideally, a fitting procedure should give parameters, the error estimates on the parameters and a statistical measure of goodness-of-fit.

3.2 Modeling the National Asthma Hospitalization Data

As noted previously, the most dramatic development in the National Asthma Hospitalization Data is the appearance of the September peak. A nonlinear model was fitted to the counts for weeks 30 to 42. The function chosen to model the observational data can be described as a normal distribution curve superimposed on a linear and quadratic background. The specific model was

$$y = a_0 + a_1 x + a_2 x^2 + a_3 \exp\left\{-a_4 (x - a_5)^2\right\}$$

where y is the count and x is the week. The model contains five parameters where a_5 represents the location of the peak in the distribution. We focus on a_5 as the main parameter of interest. The nonlinear model was fitted using the Levenberg-Marquardt algorithm for non-linear least squares in a computer program called Slide-Write (Advanced Graphics Software). This algorithm starts with initial trial values from the original data and improves the solution by minimizing the differences between observed and expected values using the gradient of the chi-square function. The nls function in S-Plus was also used to fit the model to the data in order to check the Slide-Write results. The primary aim of the nonlinear modelling was to provide estimates of parameter a_5, the location of the September peak. Slide-Write allowed us to calculate the time of the September peak with greater accuracy than using the raw data alone.

This method was used to model the data on children aged 2 to 4 and 5 to 15 as well as data based on gender, regions and latitude groupings in children aged 2 to15. Nonlinear least squares modelling was also done on region by age and latitude by age data. Once the peak positions were estimated, they were entered into an alternate database in order to examine the differences by region, latitude and age.

3.3 Nonlinear Least Squares Method

Nonlinear models depend nonlinearly on the set of M unknown parameters a_k, $k = 1,\ldots,M$. In order to fit a nonlinear model to the data, typically a χ^2 merit function is defined and then the best-fit parameters are determined by its minimization. With nonlinear dependences, the minimization must proceed iteratively. Trial values for the parameters are initially selected. Next, a procedure is developed which improves upon the trial solution. The procedure is repeated until χ^2 effectively stops decreasing. The procedure, as shown in Bates and Watts (1988) involves the following steps. If the approximation to the χ^2 function is sufficiently close to the minimum, then it is a good approximation. At this point, the minimizing parameters $\mathbf{a}_{\mathbf{min}}$ can be found from the current trial parameters $\mathbf{a}_{\mathbf{cur}}$ by the equation

$$\mathbf{a}_{\mathbf{min}} = \mathbf{a}_{\mathbf{cur}} + \mathbf{D}^{-1}\left(-\nabla\chi^2(a_{cur})\right) \qquad (4.4.1)$$

where $\mathbf{D}$ is the M x M second derivative matrix (Hessian matrix) of the χ^2 merit function at any $\mathbf{a}$. If the approximation to the χ^2 function is poor, then the next trial parameter $\mathbf{a}_{\mathbf{next}}$ can be found using

$$\mathbf{a}_{\mathbf{next}} = \mathbf{a}_{\mathbf{cur}} - \text{constant} \times \nabla\chi^2(\mathbf{a}_{\mathbf{cur}}) \qquad (4.4.2)$$

where the constant is small enough not to exhaust the downhill direction. In order to use equation (4.4.1), the gradient of the χ^2 function must be computable at any set of parameters **a**. Similarly, to use equation (4.4.2), the matrix **D** is needed.

The above steps for fitting a nonlinear model will now be shown in detail for the nonlinear model chosen for the study. Let's represent the nonlinear model to be fitted by $y = y(x;\mathbf{a}) = a_0 + a_1x + a_2x^2 + a_3\exp\left(-a_4(x-a_5)^2\right)$. The merit function is therefore given by

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N}\left[\frac{y_i - y(x_i;\mathbf{a})}{\sigma}\right]^2 = \sum_{i=1}^{N}\left[\frac{y_i - \left(a_o + a_1x_i + a_2x_i^2 + a_3\exp(-a_4(\ x_i - a_5)^2)\right)}{\sigma}\right]^2 \quad (4.4.3)$$

which is the usual sum of squares criterion. The gradient of χ^2 with respect to the parameters **a** has components

$$\frac{\partial\chi^2}{\partial a_k} = -2\sum_{i=1}^{N}\frac{[y_i - y(x_i;\mathbf{a})]}{\sigma^2}\frac{\partial y(x_i;\mathbf{a})}{\partial a_k} \qquad k = 1,2,\ldots,M\,. \qquad (4.4.4)$$

The gradient of χ^2 will be zero when the χ^2 function is at its minimum. Taking the second derivative results in the Hessian matrix, **D,** which is given by

$$\frac{\partial^2\chi^2}{\partial a_k\partial a_l} = 2\sum_{i=1}^{N}\frac{1}{\sigma^2}\left[\frac{\partial y(x_i;\mathbf{a})}{\partial a_k}\frac{\partial y(x_i;\mathbf{a})}{\partial a_l} - [y_i - y(x_i;\mathbf{a})]\frac{\partial^2 y(x_i;\mathbf{a})}{\partial a_l\partial a_k}\right] \qquad (4.4.5)$$

It is conventional to remove the 2 in equations (4.4.4) and (4.4.5) by defining

$$\beta_k \equiv -\frac{1}{2}\frac{\partial \chi^2}{\partial a_k} \qquad \alpha_{kl} \equiv \frac{1}{2}\frac{\partial^2 \chi^2}{\partial a_k \partial a_l} \tag{4.4.6}$$

At the χ^2 minimum, $\beta_k = 0$ for all *k*. Equation (4.4.1) can be rewritten as the set of linear equations

$$\sum_{l=1}^{M} \alpha_{kl}\delta_{al} = \beta_k \tag{4.4.7}$$

This set of linear equations is solved for the increments δ_{al} which when added to the current approximation give the next approximation. Equation (4.4.2) translates to

$$\delta_{al} = \text{constant} \times \beta_l \tag{4.4.8}$$

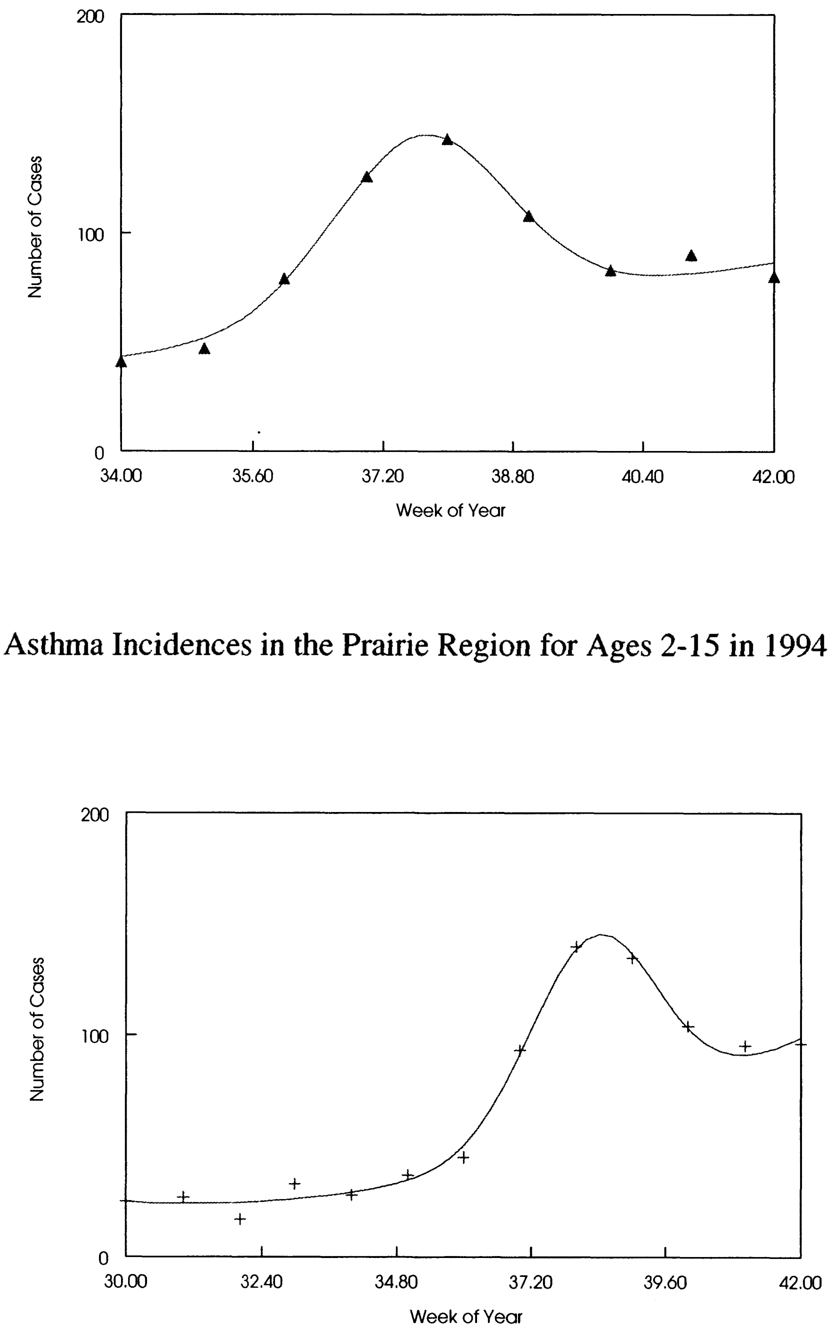
The Hessian matrix depends on both the first and second derivatives of the χ^2 merit function. The second derivative occurs since the gradient (4.4.4) has a dependence on the first derivative $\partial y / \partial a_k$. The second derivative may be ignored when it is zero or small enough when compared to the term with the first derivative. Furthermore in practice, the term multiplying the second derivative, $[y_i - y(x_i;\mathbf{a})]$, should represent the random measurement error of each data point. This error should be uncorrelated with the model. Hence, when summed over *i*, the second derivative terms usually cancel out. For these reasons, the second derivative term may be dismissed. Thus α_{kl} may be rewritten as

$$\alpha_{kl} = \sum_{i=1}^{N} \frac{1}{\sigma^2}\left[\frac{\partial y(x_i;\mathbf{a})}{\partial a_k}\frac{\partial y(x_i;\mathbf{a})}{\partial a_l}\right] \qquad (4.4.9)$$

3.4 Results of the Nonlinear Least Squares Modelling

The results of the nonlinear least squares modelling using Slide-Write were essentially identical to the results from S-Plus. The computed R^2 of all the curve fits ranged from 0.96 to 0.99 with the one exception being 0.82. This gives some assurance that the nonlinear model, $y = a_0 + a_1 x + a_2 x^2 + a_3 \exp\left\{-a_4(x - a_5)^2\right\}$ fits the data well. Figure 3.1 depicts the data used and the model fitted for two cases. Apart from a small number of cases, the results from other fits show very similar features.

Figure 3.1 Nonlinear Curve Fits for Two Cases.



a) Asthma Incidences in the Prairie Region for Ages 2-15 in 1994

b) Asthma Incidences for Latitude Group Under 44 for Ages 2-4 in 1990

Table 3.1 shows the estimates for the parameter, a_5, which is the location of the annual September peak and its standard error by age groups and by gender. Table 3.2 gives the annual September peaks and the corresponding standard errors by region. Similar tables were constructed for the location of the September peaks by latitude, latitude by age and region by age groupings. The researcher (Mr. Neil Johnston) was very interested in these peaks and the results of the nonlinear fits.

Table 3.1 Location of September Peaks by Age and Gender.

Year	*Age*		*Gender*	
	Age 2-4 (SE)	*Age 5-15 (SE)*	*Males (SE)*	*Females (SE)*
1990	38.217 (0.077)	38.038 (0.065)	38.408 (0.199)	38.309 (0.189)
1991	38.104 (0.124)	37.848 (0.091)	38.161 (0.249)	38.259 (0.275)
1992	38.512 (0.128)	38.435 (0.081)	38.817 (0.209)	38.923 (0.269)
1993	38.169 (0.111)	37.782 (0.079)	38.023 (0.209)	38.031 (0.204)
1994	37.926 (0.130)	37.834 (0.080)	38.142 (0.150)	38.121 (0.194)
1995	37.929 (0.103)	37.763 (0.055)	37.967 (0.111)	37.949 (0.151)
1996	37.834 (0.071)	37.690 (0.056)	37.966 (0.182)	37.979 (0.230)
1997	37.782 (0.182)	37.284 (0.079)	37.512 (0.113)	37.599 (0.152)
1998	38.268 (0.075)	37.013 (0.087)	38.439 (0.192)	38.536 (0.307)
1999	38.299 (0.102)	37.923 (0.057)	38.104 (0.096)	38.096 (0.129)
2000	38.316 (0.136)	38.081 (0.078)	38.649 (0.221)	38.387 (0.249)

Table 3.2 Location of September Peaks Grouped by Region.

Year	**BritishColumbia** *(SE)*	*Prairie Region (SE)*	*Ontario (SE)*	*Atlantic Region (SE)*
1990	38.177 (0.084)	37.943 (0.121)	38.158 (0.072)	38.062 (0.131)
1991	37.781 (0.155)	37.464 (0.221)	38.322 (0.051)	37.957 (0.109)
1992	38.634 (0.119)	38.034 (0.408)	38.590 (0.064)	38.276 (0.262)
1993	38.255 (0.158)	37.350 (0.556)	37.901 (0.094)	37.700 (0.163)
1994	37.804 (0.103)	37.679 (0.067)	37.939 (0.121)	37.890 (0.128)
1995	37.842 (0.694)	37.769 (0.075)	37.809 (0.065)	37.949 (0.327)
1996	37.501 (0.045)	37.348 (0.132)	37.958 (0.036)	37.601 (0.119)
1997	37.958 (0.194)	37.174 (0.074)	37.335 (0.102)	37.255 (0.145)
1998	38.319 (0.152)	37.976 (0.158)	38.390 (0.081)	37.518 (0.098)
1999	38.304 (0.123)	37.511 (0.068)	38.167 (0.080)	38.516 (0.104)
2000	38.165 (0.077)	37.839 (0.224)	38.443 (0.041	37.595 (0.206)

Chapter 4: Likelihood Ratio Test for Homogeneity of the September Peak Locations

4.1 Likelihood Ratio Test

An issue of interest was to establish the consistency of the September peak locations across the various groups. After some thought, we decided to use a likelihood ratio approach. The likelihood ratio test provides the means for comparing the likelihood of the data under one hypothesis called the alternative hypothesis against the likelihood of the data under another, more restricted hypothesis, called the null hypothesis. The likelihood ratio test measures how likely is that the data come from the null hypothesis relative to the alternative hypothesis.

The basic idea as explained by Hogg and Craig (1995, pp. 409-422) is to compare the maximized likelihoods under the two hypotheses. The maximized likelihood under the null hypothesis, H_o, is $\max_{\theta \in H_o} L(\theta, y) = L(\hat{\theta}_{H_o}, y)$ where $\hat{\theta}_{H_o}$ denotes the mle of θ under the null hypothesis. The maximized likelihood under the larger, alternative hypothesis, H_1, is $\max_{\theta \in H_1} L(\theta, y) = L(\hat{\theta}_{H_1}, y)$ where $\hat{\theta}_{H_1}$ denotes the mle of θ under the alternative hypothesis. The ratio of these two quantities, $\lambda = \frac{L(\hat{\theta}_{H_o}, y)}{L(\hat{\theta}_{H_1}, y)}$, is bound to be between 0 and 1 since likelihoods are non-negative and the likelihood of the restricted model can not exceed that of the larger model because it is nested on it. Values close to 0 indicate that the smaller model is not acceptable compared to the larger model, because it would make the observed data very unlikely. Values close to 1 indicate that the smaller model is almost as good as the large model, making the data just as likely. Under certain regularity conditions, minus twice the log of the likelihood ratio has approximately, in large samples, a chi-square distribution with degrees of freedom equal to the difference in the number of parameters under the two hypotheses. Thus,

$$\Lambda = -2\log\lambda = 2\log L\left(\hat{\theta}_{H_1}, y\right) - 2\log L\left(\hat{\theta}_{H_o}, y\right)$$

is approximately distributed as χ_v^2, where v is the degrees of freedom.

4.2 Application of Likelihood Ratio Test to the September Peaks

Denote by $\hat{a}_{ij}$ the estimated position of the September peak for year j in the i^{th} group, $1 \le j \le n$, $1 \le i \le m$. In our case, $n = 11$ years and $m =$ 2, 4 or 8 groups. Our analysis is based on the assumption of independence and that $\hat{a}_{ij}$ is normally distributed with some underlying mean location a_i and standard deviation σ_{ij} identical to the estimated value $\hat{\sigma}_{ij}$ from Chapter 3. In this context, we aim to test

$$H_o : a_1 = a_2 = \cdots = a_m \text{ vs. } H_1 : H_o \text{ is false.}$$

Thus, H_o is the hypothesis that the underlying location of the September peaks are the same for all the groups.

The likelihood function under H_1 is

$$L(a_1, a_2, \ldots, a_m) = \prod_{i=1}^{m}\prod_{j=1}^{n}\frac{1}{\sqrt{2\Pi}\sigma_{ij}}\exp\left[-\left(\frac{(\hat{a}_{ij}-a_i)^2}{2\sigma_{ij}^2}\right)\right] \propto \prod_{i=1}^{n}\prod_{j=1}^{m}\exp\left[-\left(\frac{(\hat{a}_{ij}-a_i)^2}{2\sigma_{ij}^2}\right)\right] = \prod_{i=1}^{m}\exp\left[-\sum_{j=1}^{n}\frac{(\hat{a}_{ij}-a_i)^2}{2\sigma_{ij}^2}\right]$$

yielding the log-likelihood $l(a_1, a_2, \ldots, a_m) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(\hat{a}_{ij}-a_i)^2}{2\sigma_{ij}^2}$. Thus the unrestricted

MLE's of $a_1, a_2, \ldots, a_m$ are obtained by solving

$$\frac{\partial l}{\partial a_i} = -\sum_{j=1}^{n}\frac{2(\hat{a}_{ij}-a_i)}{2\sigma_{ij}^2}(-1) = 0$$

yielding

$$\hat{a}_i = \frac{\sum_{j=1}^{n}\frac{\hat{a}_{ij}}{\sigma_{ij}^2}}{\sum_{j=1}^{n}\frac{1}{\sigma_{ij}^2}} \quad \text{for } 1 \le i \le m.$$

The log-likelihood ratio under the null hypothesis H_o that $a_1 = a_2 = \cdots = a_m = a$ is

$$l(a) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(\hat{a}_{ij}-a)^2}{2\sigma_{ij}^2}.$$

The restricted MLE's of $a_1, a_2, \ldots, a_m$ are obtained by solving for a in

$$\frac{\partial l}{\partial a} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(\hat{a}_{ij}-a)}{\sigma_{ij}^2} = 0$$

where $\tilde{a}_1 = \tilde{a}_2 = \cdots = \tilde{a}_m = \hat{a}$. Thus solving for $\hat{a}$, the restricted MLE's of $a_1, a_2, \ldots, a_m$ is

$$\tilde{a}_i = \hat{a} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\hat{a}_{ij}}{\sigma_{ij}^2}}{\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{\sigma_{ij}^2}} \quad 1 \le i \le m.$$

The likelihood ratio test statistic is then

$$\Lambda = -2\ln\frac{L(\tilde{a}_1,\tilde{a}_2,\ldots,\tilde{a}_m)}{L(\hat{a}_1,\hat{a}_2,\ldots,\hat{a}_m)} = -2\ln\frac{\prod_{i=1}^{m}\exp\left(-\sum_{j=1}^{n}\frac{(\hat{a}_{ij}-\hat{a})^2}{2\sigma_{ij}^2}\right)}{\prod_{i=1}^{m}\exp\left(-\sum_{j=1}^{n}\frac{(\hat{a}_{ij}-\hat{a}_i)^2}{2\sigma_{ij}^2}\right)}$$

that is

$$\Lambda = -2\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{-(\hat{a}_{ij}-\hat{a})^2}{2\sigma_{ij}^2}+\frac{(\hat{a}_{ij}-\hat{a}_i)^2}{2\sigma_{ij}^2}\right)\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(\hat{a}_{ij}-\hat{a})^2-(\hat{a}_{ij}-\hat{a}_i)^2}{\sigma_{ij}^2}.$$

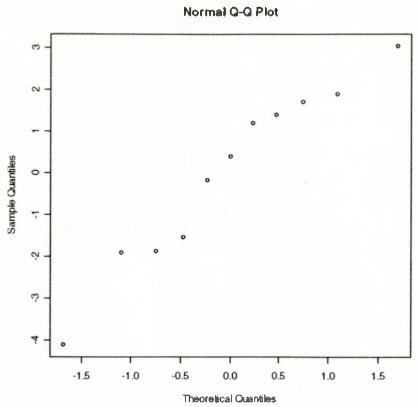
It has an approximate chi-square distribution with $(m-1)$ degrees of freedom.

We first made Normal Q-Q plots of the estimates for each group to assess the assumption of normality. Figure 4.1 shows some of the plots. The linear trend was evident for nearly all of the Q-Q plots thus supporting the assumption of normality. Specifically we plotted the standardized quantities

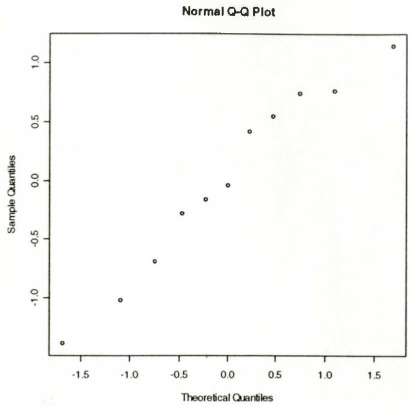
$$\hat{e}_{ij} = \frac{\hat{a}_{ij}-\hat{a}_i}{\sigma_{ij}}, \quad 1 \le j \le n.$$

If the normality assumption is correct, the $\hat{e}_{ij}$ should be roughly standard normally distributed. The estimates and standard errors came from Table 3.1 and 3.2.

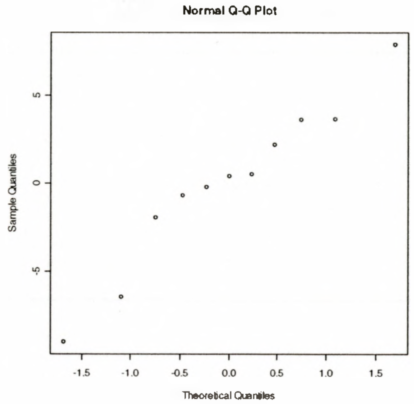
Figure 4.1 Normal Q-Q Plots



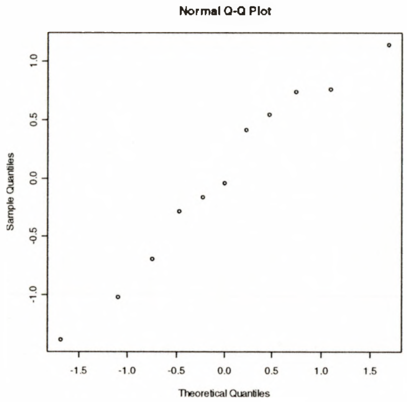
Age 2 to 4



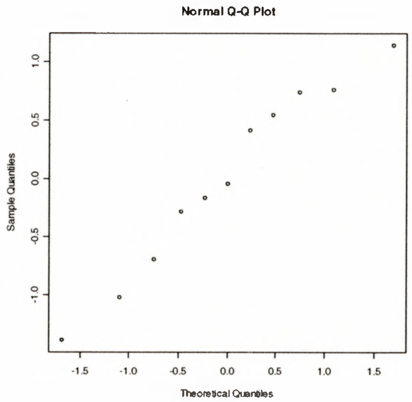
Prairie



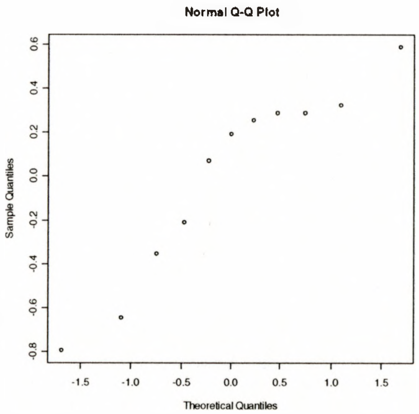
Age 5 to 15



Males



Latitude 44-50



Aged 2 to 4 Latitude Under 44

4.3 Likelihood Ratio Test Results

S-Plus was used to carry out the likelihood ratio test for the factors, age, region, region by age, latitude, latitude by age and gender. The S-Plus function given in Appendix B was written to make the required calculations.

The results (taking $\alpha = 0.05$) of the likelihood ratio tests are shown in Table 4.1. Note that, since the underlying distribution is normal, the chi-squared distribution for the likelihood ratio test is exact in this case. It can be seen that the factor age gave a significant result. This suggests that the positions of the annual asthma hospitalization peak are significantly different for the two age groups. Similarly, the positions of the annual asthma hospitalization peaks were found to be significantly different between the four regions and between the four latitudes. In addition, the positions of the annual asthma peak were significantly different for the crossing between region and age and for latitude and age. Gender was the only factor with a nonsignificant likelihood ratio test result, that is, the annual asthma peak positions were not significantly different between males and females.

Table 4.1. Likelihood Ratio Test Results.

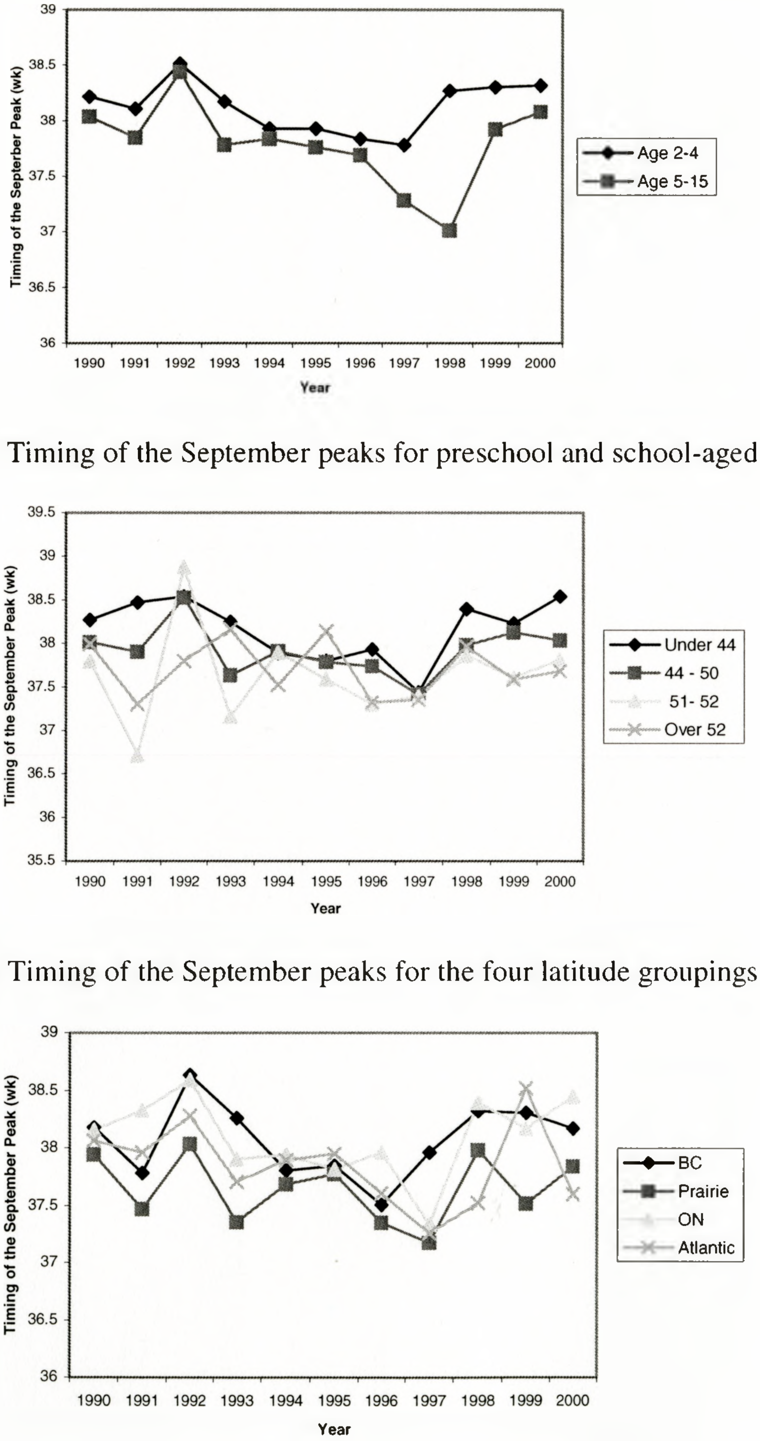
Factor	Levels	Likelihood Ratio Statistic	P-value
Age	Age 2-4 and Age 5-15	72.25	< 0.0001
Region	BC, Prairie, ON, Atlantic	283.3	< 0.0001
Region by Age	BC Aged 2-4, BC Aged 5-15 Prairie Aged 2-4, Prairie Aged 5-15 ON Aged 2-4, ON Aged 5-15 Atlantic Aged 2-4, Atlantic Aged 5-15	478.3	< 0.0001
Gender	Male and Female	0.0754	0.7836
Latitude	Under 44, 44-50, 51-52, Over 52	197.3	< 0.0001
Latitude by Age	Under 44 Aged 2-4, Under 44 Aged 5-15 44-50 Aged 2-4, 44-50 Aged 5-15 51-52 Aged 2-4, 51-52 Aged 5-15 Over 52 Aged 2-4, Over 52 Aged 5-15	373.1	< 0.0001

The timing of the September peaks for preschool and school-aged children appears to move in unison as shown in Figure 4.2. However, there was a significant difference between the two distributions with the preschool children's peaks occurring an average of 2.1 days later than the peaks of school-aged children. The timing of the September peak for the preschool children occurred after that of the school-aged children's during every year of study period. The range for the occurrence of the September peaks was between 0.5 to 7.5 days. In fact, the timing of the peak for preschool children occurs later than the school-aged children's peak for every region and latitude range in the study period.

There also appears to be a linear gradient in the timing of the September peak from the northernmost (earliest) to the southernmost (latest) range of latitude shown in Figure 4.2. The difference from north to south was approximately 2.8 days.

In addition Figure 4.2 also shows the timing of the September peaks in the four regions over the eleven years. Although the occurrence of the peaks was found to differ significantly across the four regions of Canada, no obvious trend in the differences was detected. The timing of the peak was earlier in the prairie and Atlantic provinces than in Ontario and British Columbia.

Figure 4.2 Timing of the Peaks by Age Group, Latitude and Region.



a) Timing of the September peaks for preschool and school-aged children

b) Timing of the September peaks for the four latitude groupings

c) Timing of the September peaks for the four regions

Chapter 5: Poisson Log-Linear Modelling of the Pediatric Asthma Hospitalization Data

5.1 Application of Poisson Log-Linear Modelling to Asthma Hospiatlizaion Data

Our analyses in previous chapters, particularly Chapter 2, focused on assessing the effect that each individual factor (age, gender, latitude, region) had on the counts of asthma cases. These analyses were marginal in that the factors were looked at one at a time. Also no time effect was considered since the counts were aggregated over the 10 years analyzed. In this Chapter we aim to alleviate these deficiencies by assessing the factor effects jointly and by incorporating time as another factor. Thus we wish to model the relationship between the count of pediatric asthma hospitalizations, y , and the explanatory variables, gender, age, latitude, risk group size and year. The response variable, y , is assumed to vary independently across the sets of conditions and follows a Poisson distribution with mean μ. Naturally μ will vary across the sets of conditions. The range of y encompasses all non-negative integers. The log density may be written as

$$\log f(y;\mu) = \log P(Y = y) = y \log \mu - \mu \ - \log y! \ .$$

The explanatory variables, $\mathbf{x} = (x_1, x_2, \ldots, x_m)'$, are used to explain the variation in the response, y, through a set of m unknown regression parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)'$. We assume a linear relationship

$$\eta = \sum_{j=1}^{m} \beta_j x_j = \boldsymbol{\beta}'\mathbf{x}$$

where η is an appropriate function of μ. If x_j is the value of a quantitative covariate then β_j scales x_j to give its effect on η. If x_j represents the presence or absence of a level of a factor then β_j is the effect of that factor level. The mean μ of the y, is related to the linear predictor η through the link function $g(\mu \ )$ which in our case is given by $\log \mu$, that is $\eta = g(\mu) = \log \mu$. The method of maximum likelihood is used to estimate the linear parameters $\boldsymbol{\beta}$ and hence the linear predictors and fitted values $\hat{\mu}$. Here, the likelihood L is a function of $\boldsymbol{\beta}$. In general, we may find the maximum of L by finding the maximum of log L. The maximum likelihood estimates may be found by solving

$$\frac{\partial \log L}{\partial \beta_j} = 0, \text{ for } j = 1, \ldots, m.$$

An iterative numerical algorithm such as the Newton-Raphson approach is used to achieve the results. For more details refer to Cameron and Trivedi (1998).

The linear structure models the effect of explanatory variables on the response variable. The data give information on which effects have an important influence and which may be neglected. A smaller number of parameters means easier interpretation and generally, better prediction. Our aim is to find the best "trade-off" between the number of parameters that must be included in the linear structure and the ability of the model to represent the data.

In order to determine the usefulness of adding parameters to a given model or the lack of fit from omitting those parameters a goodness of fit measure is needed. A widely used method of goodness of fit compares the likelihood of the current model, L_c with the likelihood of the full model L_f. This method is based on the ratio L_c/L_f. It is more convenient to find the scaled deviance, $S(c,f)$ which is defined as

$$S(c,f) = -2\log(L_c/L_f)$$

Large values of $S(c,f)$ reveal lack of fit of the given model.

Suppose L_1 is the likelihood for model 1 and L_2 is the likelihood for model 2. Assuming that model 2 is nested in model 1 then if model 2 is correct, the scaled deviance $S(2,1) = -2\log(L_2/L_1)$ is approximately distributed as χ^2 with $t_1 - t_2$ degrees of freedom where t_i is the number of independent parameters under model *i*. By analyzing the scaled deviance, it is possible to assess the usefulness of the parameters in the general linear model. The scaled deviance for a Poisson model is given by

$$S = 2\sum_i \left\{ y_i \log\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\}$$

with $n-t$ degrees of freedom where n is the number of observations and t is the number of parameters in the model of interest. For more details refer to Draper and Smith (1998, pp. 600-610) and Dobson (2001, pp. 76-80).

The response and explanatory variables considered for the analysis are described in Table 5.1. The covariate, Size, was found by adjusting the 1996 Census data for population growth for the years 1991 to 2000. The population growth rate was found using the 2003 Canadian Almanac. Given the groupings considered in Table 5.1, a total of 160 group counts were created and analyzed.

Table 5.1 Variables in the Dataset used in the Poisson Log-Linear Modelling.

Variable	**Description**
Count (C)	Number of asthma hospitalization case (numerical)
Age (A)	Age 2 to 4 or Age 5-15 (2 groups)
Gender (G)	Males or Females (2 groups)
Latitude (L)	Under 44, 44-50, 51-52 or Over 52 (4 groups)
Size (S)	Population size of particular variable grouping (numerical)
Year (Y)	1991-2000 (numerical)

5.2 Results of the General Linear Modelling

A good strategy for dealing with a complex situation is to begin with a model that includes all the second order interactions. If the second order interactions appear negligible then the usual sequence of progressively simplified models can start from there. However, if there appears to be more complex relationships then the data can be split into subgroups. This strategy allows for a well-structured analysis without too many assumptions. Another method is to use backward or forward elimination to select the best model.

Starting with the model

$$\ln \mu_{ijklm} = \beta_0 + \beta_i^A + \beta_j^G + \beta_k^L + \beta_l^Y + \beta_m^S + \beta_{ij}^{AG} + \beta_{ik}^{AL} + \beta_{jk}^{GL},$$

all the two-way interactions were found to be significant. Successively simplifying the model did not improve the fit since eliminating parameters from the model increased the scaled deviance significantly. Furthermore, adding successively complex interactions to the model did not improve the fit enough to warrant keeping them in the model. Using model selection techniques also returned the model with all the two-way interactions as the best choice.

The scaled deviance of the final model was 987.37 on 145 degrees of freedom. Dividing the residual deviance by its degrees of freedom gives a value of 6.81 which is significantly larger than 1, giving evidence of overdispersion. This technique is used to detect overdispersion or underdispersion in Poisson regression. For a Poisson distribution the mean and the variance are equal which implies that the ratio of the deviance and its degrees of freedom should be approximately 1. Values greater than 1 indicate overdispersion, that is, the true variance is greater than the mean. Values less than 1 indicate underdispersion. Evidence of overdispersion, as in our case, indicates inadequate fit of our Poisson model. Overdispersion implies that there may be other factors needed in the model to describe the data using a Poisson distribution that we are unaware of. For details on overdispersion refer to Dean (1989). Figure 5.1 provides further evidence of overdispersion. This is a normal Q-Q plot of the transformed residuals where the transformation is done to approximate normality assuming the correct model has been fitted. (e.g. see Francis, Green and Payne, 1993, pp. 283-286). There were many residual data points in Figure 5.1 with high values that were not well explained by the model. The large residual values were examined in more detail to determine whether there were any similarities between them. No obvious trend or similarities were found.

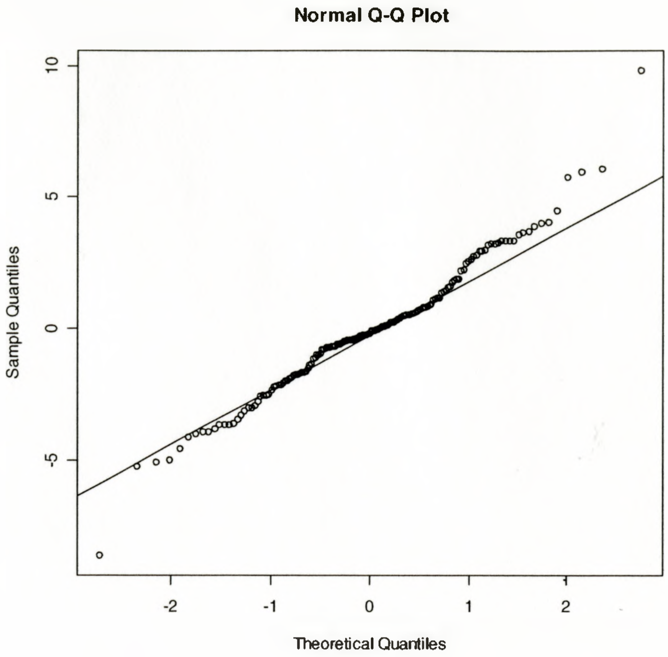


Figure 5.1 Normal Q-Q Plot of Residuals for the Final Poisson Model.

In order to account for the overdispersion, the data were refitted using a Quasi-Poisson model which allowed the dispersion parameter to be a constant other than 1. The dispersion parameter which best described the data was 6.84. The best Quasi-Poisson model was found to be

$$\ln \mu_{ijkl} = \beta_0 + \beta_i^A + \beta_j^G + \beta_k^L + \beta_l^Y + \beta + \beta_{ij}^{AG} + \beta_{ik}^{AL} + \beta_{jk}^{GL} .$$

Notice that the covariate, size, was not found to provide a significant contribution to the model and thus was eliminated. Figure 5.2 shows a plot of the fitted values against the observed values for the Quasi-Poisson log-linear model. There is a definite trend to linearity with a bit of noise.

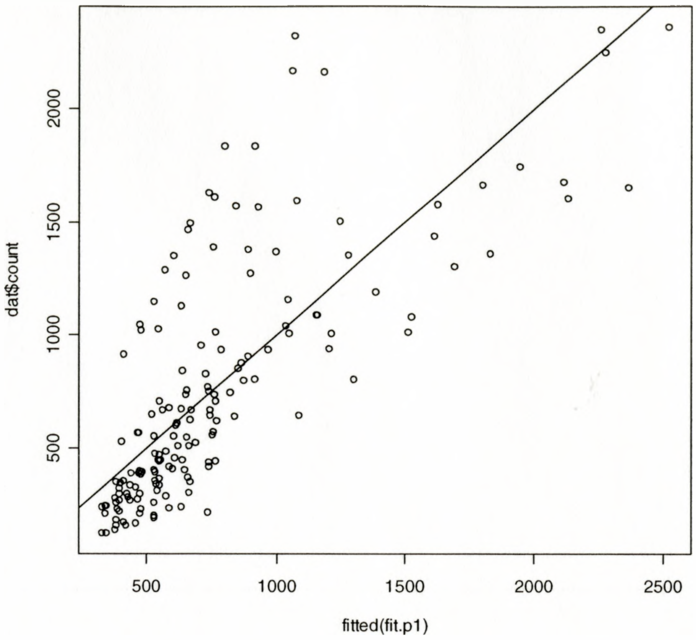


Figure 5.2 Plot of Fitted values versus Responses for the Final Quasi-Poisson
Model.

Table 5.2 gives the parameter estimates of the final Quasi-Poisson model with dispersion parameter 6.84. The average number of asthma hospitalizations for children aged 5-15 and males were found to be greater than that of children aged 2-4 and females respectively. The average number of asthma hospitalizations in latitude group 44-50 was greater than the counts for latitude group under 44 whereas the asthma hospitalization counts for latitude groups 51-52 and over 52 were both less than the counts for latitude group under 44. The negative parameter estimate for year indicates a small negative trend over the years or more specifically, decreasing asthma hospitalization admissions over the years. This may reflect a true decrease in asthma attacks or underreporting due to external factors, for instance the hospital closures that occurred in the 90's. Age was found be related to gender. Males aged 5 to 15 on average were committed to hospitals for asthma less than females aged 2 to 4. There was also a definite positive relationship found between age and latitude groups. The relation between gender and latitude groups was found to be significant only for males in latitude group over 52. Males in latitude group over 52 were less likely than females in latitude group under 44 to be admitted into hospitals for asthma.

Table 5.2 Parameter Estimates for Quasi-Poisson Model

$\ln \mu_{ijkl} = \beta_0 + \beta_i^A + \beta_j^G + \beta_k^L + \beta_l^Y + \beta + \beta_{ij}^{AG} + \beta_{ik}^{AL} + \beta_{jk}^{GL}$.

Factor	Contrast	Parameter Estimate	P-value
Age	Age 5-15 vs. Age 2-4	0.11352	0.0008 *
Gender	Males vs. Females	0.70091	< 0.0001 *
Latitude	44-50 vs. Under 44	0.23193	< 0.0001 *
	51-52 vs. Under 44	-1.28699	< 0.0001 *
	Over 52 vs. Under 44	-0.96878	< 0.0001 *
Year		-0.08545	< 0.0001 *
Age & Gender	Age 5-15 & Males vs. Age 2-4 & Females	-0.28707	< 0.0001 *
Age & Latitude	Age 5-15 & 44-50 vs. Age 2-4 & Under 44	0.21526	< 0.0001 *
	Age 5-15 & 51-52 vs. Age 2-4 & Under 44	0.42178	< 0.0001 *
	Age 5-15 & Over 52 vs. Age 2-4 & Under 44	0.32583	< 0.0001 *
Gender & Latitude	Males & 44-50 vs. Females & Under 44	-0.04373	0.2273
	Males & 51-52 vs. Females & Under 44	-0.09531	0.08445
	Males & Over 52 vs. Females & Under 44	-0.12077	0.0171*

* Significant at α level of 0.05

S-Plus and SPSS were used for the Poisson Log-Linear modeling. The results from both programs agreed on the basic model (no overdispersion modeled). However, only S-Plus allowed regression under overdispersion.

Chapter 6: Conclusions and Discussion

6.1 Discussion of Study Results

Asthma is still a major cause of morbidity despite advances in understanding the disease and more effective medications. In Canada, approximately 500 people die each year from asthma. Hospitalization of children for asthma in Canada has been shown to be highly predictable and consistent with an annual cycle that peaks in September.

The weekly distributions of asthma hospitalization counts were examined for homogeneity across age, latitudes, regions and gender over three years. We found no significant difference between the distributions of asthma cases for males and females. However, there was significant evidence that the weekly distributions of asthma counts differed between the four regions and four latitude groups over the three years. The distribution of asthma cases for children aged 2 to 4 and 5 to 15 was also significantly different but only for two out of the three years considered.

Next, a nonlinear model was fitted to the asthma hospitalization counts for weeks 30 to 42. Our main goal was to provide the parameter estimates and standard errors for the timing of the peak. Once obtained, the timing of the

September peaks was tested to establish consistency across the various factors using the Likelihood Ratio Test. Gender did not significantly affect the annual timing of the September peaks. The timing of the peaks were significantly different between the four regions and four latitude groups. In addition, the cross analyses of region by age and latitude by age were found to have significantly different peak locations

We also found a significant difference in the timing of the September peaks between preschool and school-age children. On average, the preschool children's peak occurs 2.1 days later than the peak of school-aged children. The timing of the September peak for preschool children occurred later than that of school-aged children for every region and latitude range during the study period.

Mr. Neil Johnston suggested the following explanation as a possible cause for the September peak. The start of school in September may provide an opportunity for children to be exposed to factors that exacerbate asthma. Children are exposed to changes in allergen exposure on their return to school. For instance, cat allergen on clothing may be transported to school by children. Fungal and pollen allergens may also be more prevalent and intense at school. Asthmatic children are frequently infected from viruses. Studies have examined the relation between viral infections and school. An increase in asthma hospitalization was found after school vacations or breaks, especially in September over eleven years (Storr and Lenney 1989). This suggests that the trends are consistent with breaks in viral transmission during vacations.

In Canada, the annual asthma hospitalization cycle for children aged 2 to 15 undergoes a significant increase in asthma hospitalizations before Labour Day. This suggests that the factors which exacerbate asthma have already been increasing before School return. Thus exposure of children to allergens and infections from school return is a plausible explanations for some cases in the September peak.

We found that the September peak in preschool children occurs approximately 2 days later than that of school-aged children. Preschool children may have older siblings who expose them to infections and allergens from school. In addition some preschoolers may attend early education programs that also increase their exposure to infections and allergens. The occurrence of the September peak in preschool children after that of the school-aged children is consistent with exposure to agents that exacerbate asthma following the return to school of older siblings.

A Poisson log-linear model was used to describe the relationship between the count of pediatric asthma hospitalizations and the explanatory variables, gender, age, latitude, risk group size and year. The best model was found to be a Quasi-Poisson log-linear model with dispersion parameter 6.84. Gender, age, latitude group and year were found to be significant in accessing pediatric asthma hospitalization counts. The two-way interactions between age and gender, age and latitude and gender and latitude were all found to be significant.

6.2 Future Direction for Research

While school return provides an opportunity for children to be exposed to factors that exacerbate asthma, the contribution of each factor and their interactions require further research. In addition, the role of older siblings as agents of exacerbation factors in young children needs to be explored. In order to research the role of older siblings, hospital records from members of the same family need to be linked.

Our analyses focused on incidences of asthma attacks without regard to whether they came from the same or different subjects. The database contains subject identifiers but that information was not given to us. It will be of interest to study the incidences distinguishing patients. A Poisson process could be used to model the times of incidences for each patient and methods for survival analysis could be applied to quantify and compare incidence rates.

6.3 Computing Issues

The grouped data came in Excel format from which S-Plus and SPSS versions were made for the statistical analyses. The raw data came in D-base format. In raw form, the data could not be use in Excel or GLIM due to its large size. Instead, S-Plus and SPSS were used for the analysis involving the raw data.

A total of 308 nonlinear least squares curve fits were conducted for Chapter 3 out of which six cases or 2% gave inappropriate results. Using Slide-Write and S-Plus, the curve fits of four cases from the region by age grouping and two cases from the latitude by age grouping gave local maximum or minimum instead of the overall global maximum for the a_5 (timing of the peak parameter). Each case was examined individually and an alternate interval for the analysis was determined instead of using weeks 30 to 42.

The histograms in Chapter 1 were done in Excel. All other plots were done in S-Plus and SPSS.

Appendix A

S-Plus function used to calculate the Chi-square test statistic in the contingency table analyses. The argument n_{ij} represents the asthma count for the specific level of the factor.

```
> contab1
function(nij)
{

# nij is the value of observed counts

        n <- sum(nij)
        cc <- dim(nij)[2]
        rr <- dim(nij)[1]
        eij  <-  (apply(nij,  1,  sum)  %o%  apply(nij,  2,
sum))/n
        chi <- sum(((nij - eij)^2)/eij)
        df <- (rr - 1) * (cc - 1)
        pv <- (1 - (pchisq(chi, df)))
        print(cc)
        print(rr)
        print(n)
        print(df)
        print("Chi-square statistic and P-value:")
        print(c(chi, pv), digits = 4)
}
```

Appendix B

S-Plus function for calculating the likelihood ratio test statistic.

```
LRatio
function(est, sd)
{
# est and sd are the values of the peak locations and their
corresponding standard errors

        m <- dim(est)[2]
        n <- dim(est)[1]
        num1 <- apply(est/sd^2, 2, sum)
        print(num1)
        den1 <- apply(sd^(-2), 2, sum)
        print(den1)
        aihat <- num1/den1
        print(aihat)
        num2 <- sum(est/sd^2)
        print(num2)
        den2 <- sum(sd^(-2))
        print(den2)
        ahat <- num2/den2
        print(ahat)
        aihatm <- matrix(c(aihat), n, m, byrow = T)
        lrstat <- sum(((est - ahat)^2 - (est -
        aihatm)^2)/sd^2)
        pvalue <- (1 - (pchisq(lrstat, m - 1)))
        print("LR Statistic and P-value:")
        print(c(lrstat, pvalue), digits = 4)
}
```

References

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc., pp. 8-13.

Bates, D. M. and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications*, NewYork: John Wiley & Sons, Inc.

Cameron, A. C. and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge: Cambridge University Press.

Dean, C. B. (1989), "Tests for detecting overdispersion in Poisson regression models". *Journal of American Statistical Association*, **84,** pp. 467-472.

Dennis, J. E. and Schnabel, R. B. (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Englewood Cliffs, NJ: Prentice-Hall.

Dobson, A. J. (2001), *An Introduction to Generalized Linear Models*, New York: Chapman & Hall/CRC.

Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis, Third Edition*, New York: John Wiley & Sons, Inc.

Francis, B., Green, M. and Payne, C. (1993), *GLIM 4 The Statistical System for Generalized Linear Interactive Modelling*, New York: Oxford University Press Inc.

Health Canada (2001), "The Prevention and Management of Asthma in Canada", The National Asthma Control Task Force.

Hogg, R. V. and Craig, A. T., (1995), *Introduction to Mathematical Statistics*, Upper Saddle River, New Jersey: Prentice-Hall Inc.

Johnston N. W., Sears M. (2001), "A national evaluation of geographic and temporal patternsof hospitalization of children for asthma in Canada [abstract]". *American Journal of Respiratory Critical Care Medicine*, **163**, A359

Lawson, C. L. and Hanson, R. (1974), *Solving Least Squares Problems*, Englewood Cliffs, NJ: Prentice-Hall.

Lindgren, B. W. (1993), *Statistical Theory*, New York: Chapman & Hall.

Storr, J., Lenney, W. (1989), “School holidays and admissions with asthma”. *Archives of Disease in Childhood*, **64**, pp. 103-107.