STATISTICAL ANALYSIS OF HUMAN PARASITIC INFECTIONS

M.Sc.

VASILEIOS V.N. NIKOLAOU

JANUARY 2001

Abstract

In longitudinal studies, measurements are taken over time or space on the same individual. A wide variety of well-established procedures exist for modelling these data such as t-tests, the use of summary measures, analysis of variance and the method of maximum likelihood. The latter seems to be the most appropriate approach since its estimates are consistent and efficient for large samples and it deals directly with problems of missing data. Here, much emphasis is given to the model proposed by A. Azzalini(1994), which is not only based on the method of maximum likelihood but it also incorporates an appropriate correlation structure for the measurements on a single individual across time. This model is applied in a real data example. It concerns a cohort study of about 1,100 individuals from China and Nigeria, where an infectious disease is widely spread. Treatment is given to control the prevalence and intensity of the disease. Our aim was to assess the way in which factors such as age and sex are related with the above outcomes, estimate the size of the effect of the treatment, as well as the time needed for prevalence and intensity to reach the pre-treatment level as a result of re-infection and make comparisons between two different types of treatment.

However, a major problem arises due to the high frequency of missing data. In this case, a generalized linear model is fitted, which although it does not take into account the correlation over time, its methods of inference are based on well-founded theory under the assumption of independent errors.

In addition, the use of smoothing techniques to explore trends in the data in a non parametric manner is described. This provides a means of modelling the data without making parametric assumptions.

Finally, the methodology of generalized additive models is used to explore non-linear effects in a model comparison setting. Such models provide a means of checking more formally on linearity assumptions and also provide a way of modelling the data even when these effects are non-linear but can be assumed to be smooth. However, they have to be treated with care since their methods of inference are approximate.

Acknowledgements

I would like to thank Prof. David Crompton for giving me access to the data and background information on the context, as well as Prof. Adrian Bowman for his supervision and help.

CONTENTS

SECTION ONE

Ascariasis and its public health significance

CHAPTER ONE

SECTION TWO

Repeated measures models and smoothing techniques for longitudinal data

CHAPTER TWO

CONTENTS

PAGE

SECTION THREE

Modelling the covariate effects

CHAPTER THREE

SECTION FOUR

Analysis of treatment effects

CHAPTER FOUR

CONTENTS

PAGE

SECTION ONE

Ascariasis and its public health significance

Chapter 1

Ascariasis and its public health significance

1.1 Introduction

It is known that 65% of the world's population who live in developing countries like Nigeria in Africa, Mexico in Latin America and China in Asia share only 17% of the world's income. This poverty is reflected in the appearance of their towns and cities which look dirty because of underdeveloped infrastructure services and facilities such as a sewage system, water supply, public sanitation and hygiene. Also, the fact that most of these towns are already old and they were planned and built by those who had neither money nor expertise, along with the increasing migration of people from villages to the urban areas, led to heightened transmission of various diseases. Furthermore, the ignorance even among educated persons about changing the way of life and the inability of the government to give financial support to the workers and their families in order to build modern houses, provides a suitable environment for the spreading out of intestinal parasites such as Ascaris lumbricoides.

1.2 Ascaris lumbricoides

Ascaris lumbricoides is an intestinal parasite and belongs to the family of roundworms, which are known as soil-transmitted or geohelminths. It is a considerable infectious and obstinate parasite with a direct life-history pattern and it is passed from human to human by the faeco-oral route during which cysts or eggs are discharged in human stools. Each female worm lives for about 1 to 2 years and releases over 200,000 eggs per day.

Depending on the structure and composition of the soil - the more clay in a soil the better the survival rate - the eggs of Ascaris can be found around houses, schools and public places. They are sticky and in areas of high risk of infection may be found on furniture, money, fruit and vegetables. They also exist in the air and household dust, so it is possible for them to be inhaled and swallowed.

Embryonation flourishes in shaded soil with temperatures between 28 and 32 degrees and moisture greater than 80%. Under these conditions, the first stage of larva is formed within about 10 days. After some days, the second larval stage, in which the larva can be seen coiled within the egg shell, is generated and it is believed that this is the stage most responsible for infection. When, finally, the larva is fully developed within the egg shell, it possesses great powers of survival and it is able to survive even under chemical treatments and adverse climatic conditions.

The exact survival time of Ascaris eggs has not yet been estimated. It depends on the climate and several time periods have been recorded. The maximum one was in Samarkand and it was estimated to be 14 years in the soil! In central Europe, the survival time is about 6 years but the proportion surviving decline sharply with time. In moderate climate, the majority of eggs survive for a few months but a small proportion of them can survive for several years. In tropical conditions, where most of the infections occur, the survival time of Ascaris eggs is still unknown.

At present, man is the natural host of Ascaris although adult worms have been found in the orangutan, dog, cat and sheep. Ascaris larva cross the mucosa of the small intestine and reach the liver and the lungs. After two months, the larva leaves the lungs and returns to the small intestine by way of the bronchii, trachea and oesophagus. This phase lasts about 2 weeks. After that is the final moult in which the larvae changes to the immature adult worm in the small intestine.

It has been estimated that one billion people harbor this worm and the disease that presents in them after the infection by Ascaris is known as ascariasis. Some features of this disease are described in Table 1.1.

Table 1.1: Features of ascariasis (Based on Stephenson, 1987, Table 4.1)

Life history events	**Clinical features**	**Effects on health**
a. Larval migration through liver and lungs	Pneumonitis, eosinophilia, fever, skin rash, range of allergic reactions	Decrease food intake, increase nitrogen loss
b. Adult worms in small intestine(chronic)	Abdominal pain, colic, nausea, disordered small bowel pattern, mucosal abnormalities	Decrease food intake, malabsorption, temporary lactose intolerance, increase nutrient excretion
c. Aggregations and migrations of adult worms(acute)	Biliary obstruction, hepatic abscess, intestinal obstruction, intussusception, pancreatitis, volvulus, worms in ectopic sites(ear, heart, thorax, vagina)	Life-threatening complications, often needing hospital admission: mortality rate unknown, perhaps 100,000 per annum (Pawlowski and Davis, 1989)

It has been shown that acute ascariasis is a considerable public health matter especially for children because it often requires hospital admission and surgery. There is also significant evidence that chronic ascariasis contributes to development and persistence of malnutrition and reduces the growth rate in children.

Most A.lumbricoides infections occurred in south-east Asia (about 70%) and fewer cases are recorded in Africa (about 12%) and Latin America (about 8%). In countries like Bangladesh, Burma, Indonesia, Malaysia, Philippines, Vietnam and some regions in India and China where the density of the human population is extremely high and the environmental conditions are favorable, ascariasis is more likely to flourish and persist. On the other hand, ascariasis tends to be rare in the arid and sparsely populated regions of the world such as Australian deserts, some South Pacific islands and in cold polar regions.

In order to reduce the spreading out of the disease, a range of control measures is available based on chemotherapy, sanitation and health education. Therefore, anthelminthic drugs which are contributing to rapid progress in the control of soil-transmitted helminthiases have been provided to the communities of the developing countries. There are three forms of anthelminthic drug administration in the community depending on the kind of treatment which is provided. Mass treatment can be offered to all individuals in an area of high epidemic and the cost is low since the drug is made available to any member of the community regardless of age, sex or infectious status. Targeted treatment is offered to a group of people, which is considered as a high risk group such as children of a primary school. In this way, children are treated better and also the school acts not only as a center for drug delivery and administration but also it can develop a modern health education. Finally, selective treatment is provided only to those who are heavily infected and it usually requires high technical costs because each individual has to be sought out when the dose of drug is due.

However, the advantages of the use of the anthelminthic drugs can be reversed within a year after the treatment as a result of re-infection. Thus, much consideration should be given in providing hygiene and health education to the community. Hygiene includes not only the improvement of sanitation systems such as the supply and management of water supply but also the treatment of "night soil" before its use as a fertilizer for the crops. Also, knowledge of how the disease is established and transmitted and how it can be faced, is essential. Therefore, the Japanese Organization for International Cooperation in Family Planning (JOICFP) has organized a program for nutrition and parasite control and they managed to bring ascariasis under control not only in Japan but also in other developing countries like Israel, South Korea and Taiwan. Although the cost of these measures is significantly high, especially for countries in which a small budget has to be shared among many health priorities, much consideration has been given to the control and treatment of ascariasis and other soil-transmitted helminthiases. (D. W. T. Crompton(1991)).

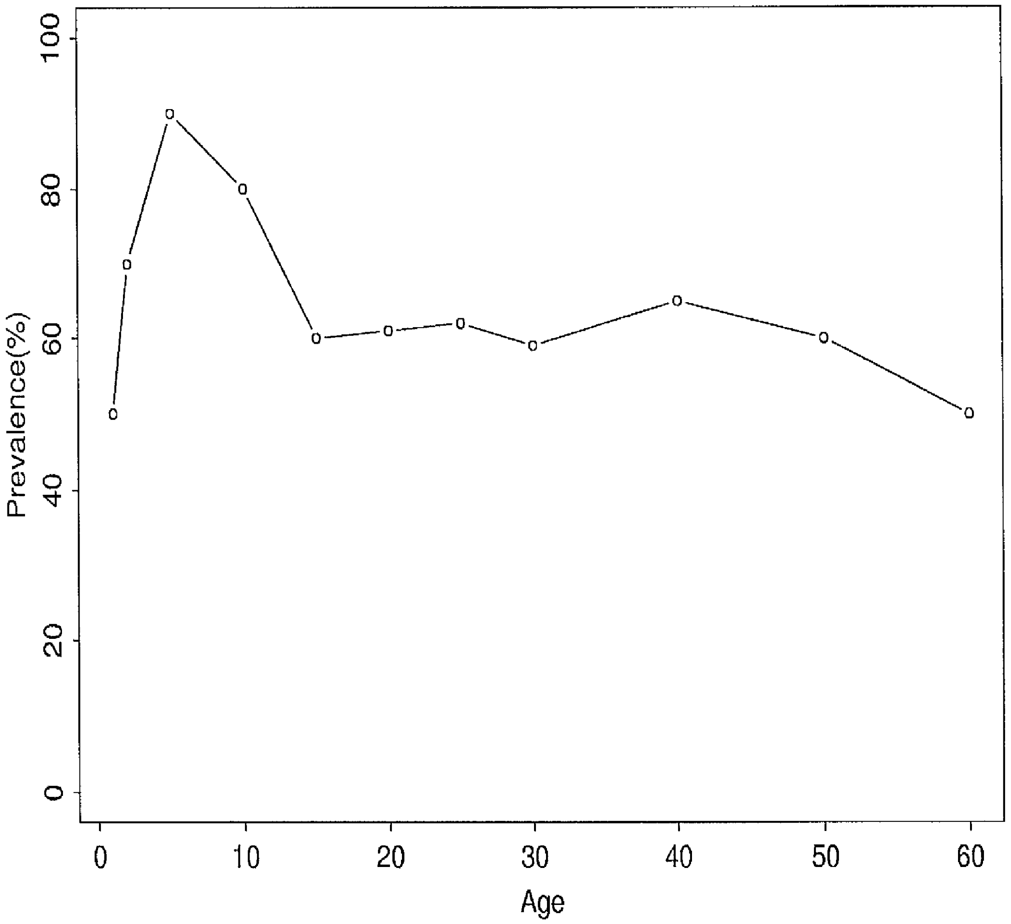
1.3 Ascaris lumbricoides infections in Asia and Africa

Recent research on ascariasis has been carried out in Nigeria, China, Ghana and other developing countries. The distribution of the disease is usually assessed by measuring the prevalence and intensity. The prevalence is determined by the number of egg-positive cases in a sample of people from the population of interest. Diagnosis is made by the detection of helminth eggs in stool samples and it is simple, quick and inexpensive. On the other hand, the intensity is defined as the number of eggs per person or per infected person and it is measured by counting either the number of worms reported in the stools or the number of eggs per gram of stool.

One of the factors which influences the prevalence of ascariasis is the environment. For example, in Africa, the prevalence is low in countries with arid climate and high in those with wet weather and warmth. Apart from the climate, the season is also a significant factor. In Saudi Arabia, the syndrome of pneumonitis that appears annually from March to May is considered to be linked to the larval migration of A.lumbricoides (Gelphi and Mustafa(1967)) associated with favorable conditions after a brief rainy season. Furthermore, housing, social status and family all have an effect on the distribution and prevalence of A.lumbricoides infection. The poorer the quality of housing the more likely it is for the parasite to flourish and persist. In addition to housing, the family influences the prevalence of ascariasis as well. When one child is found to be infected, the other members of the family are also likely to be infected.

It is common to examine the relationship between prevalence and host age because in this way it is easier to define the public health significance of the problem. A study in Manhu villages of China, which was carried out during the year June 1993 - June 1994, showed that maximum prevalence values were observed in children aged 5 to 15 years rather than in older ages (Figure 1.1). (Peng Weidong, Zhou Xianmin, Cui Xiaomin,

Figure 1.1: Prevalence of ascariasis by age in a rural community of China.

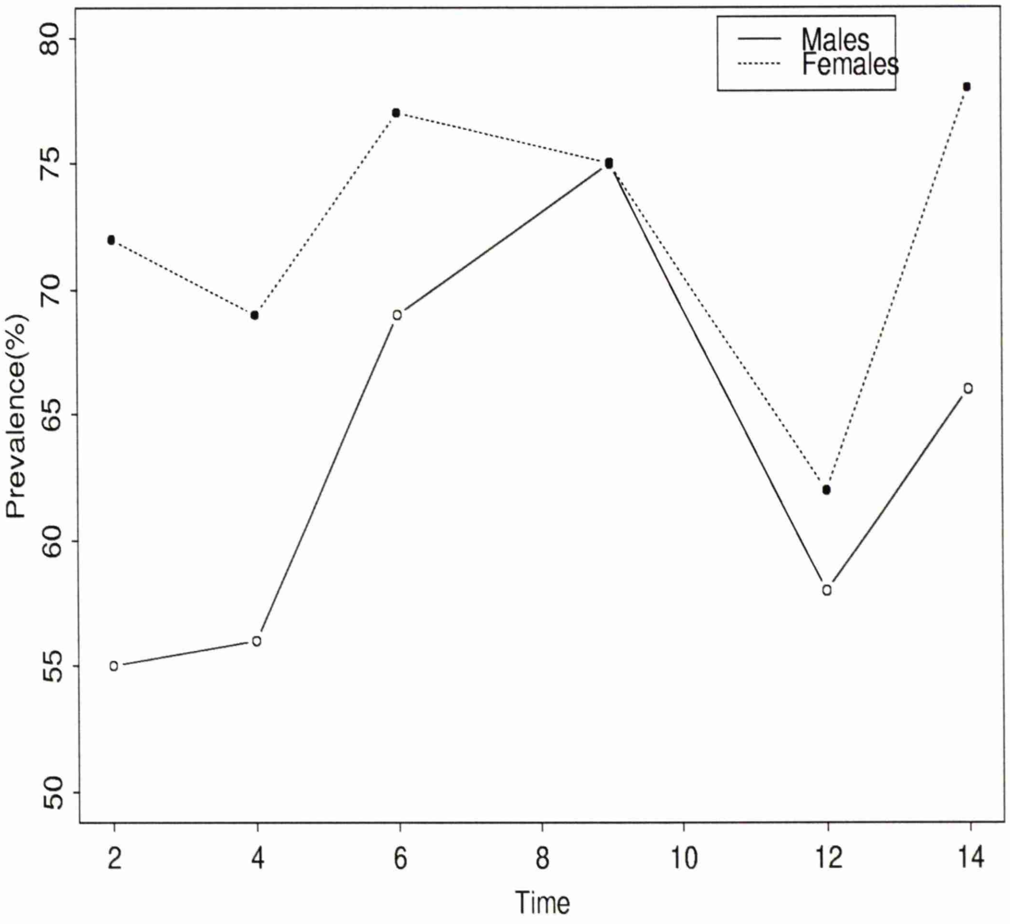


D.W.T. Crompton, R.R. Whitehead, Xiong Jiangqin, Wu Haigeng, Peng Jiyuan, Yang Yang, Wu Weixing, Xu Kaiwu and Yan Yongxing (1996)).

The main reason for this trend is that young children are particularly at risk from A.lumbricoides infection, since they are crawling and playing in contaminated soil, thus they have more possibilities of being in contact with the eggs of Ascaris. In Malaysia, the prevalence was reported to be 6% for children aged 0 to 5 years but increasing rapidly to 24.9% in children aged 5 to 15 years with a peak value of 35.6% in preschool age. It was also noted that ascariasis was most prevalent among children living in urban slums. Similar studies in Burma, Kenya and Peru showed that there is a trend for an increased prevalence in children after the first year of life and then a gradual reduction in the older age groups.

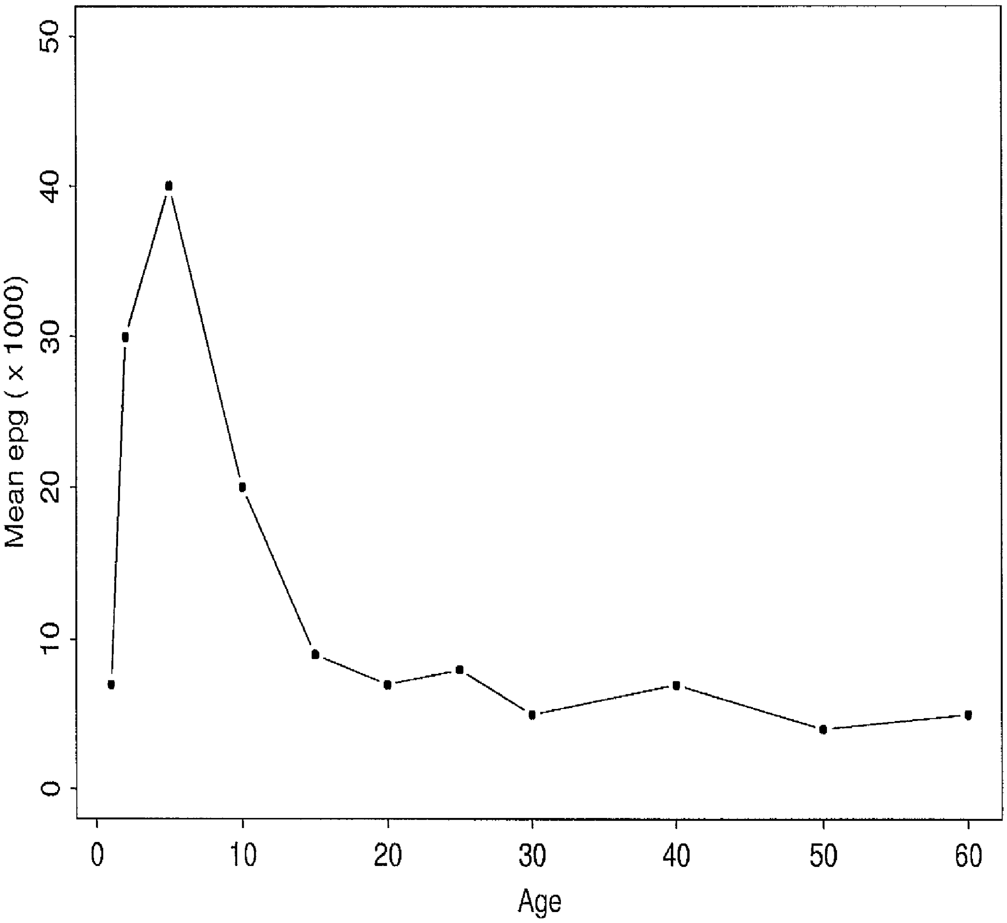
Also, the prevalence of Ascaris infection tends to vary by sex. It seems that females are more likely to be infected than males, as indicated in Figure 1.2 which plots prevalence against time(months) in Manhu villages of China.

Figure 1.2: Prevalence of ascariasis by sex in a rural community of China.



The explanation for this trend is assumed to depend on the culture of each country and usually in many of the developing countries, women are often required to work in the field and they experience greater exposure to the contaminated soil. When the relation between prevalence and host sex was examined in Kenya, it was noticed that pregnant women had a similar percentage of infection to their children (31 and 27% respectively), markedly higher than those in adult males. However, in other countries like Malaysia and Burma, there was no marked difference in the prevalence of ascariasis between boys and girls, since they have similar occupational activities and habits. Another crucial factor for understanding the population biology and transmission dynamics of Ascaris, is the intensity. As is mentioned above, intensity is the mean number of worms per person and it can be measured either by counting the number of worms passed after treatment or by counting eggs in a stool sample. Since the severity of the disease increases as the number of worms increases, one of the main purposes of many programs for the control of soil-transmitted helminthiases is to reduce the intensity of the infection. Similar to prevalence, the intensity of ascariasis(i.e. the mean number of eggs over all people in the study) in Manhu villages in China (June '93-June '94), is increasing during childhood but drops rapidly with age (Figure 1.3).

Figure 1.3: Intensity of ascariasis by age in a rural community of China.

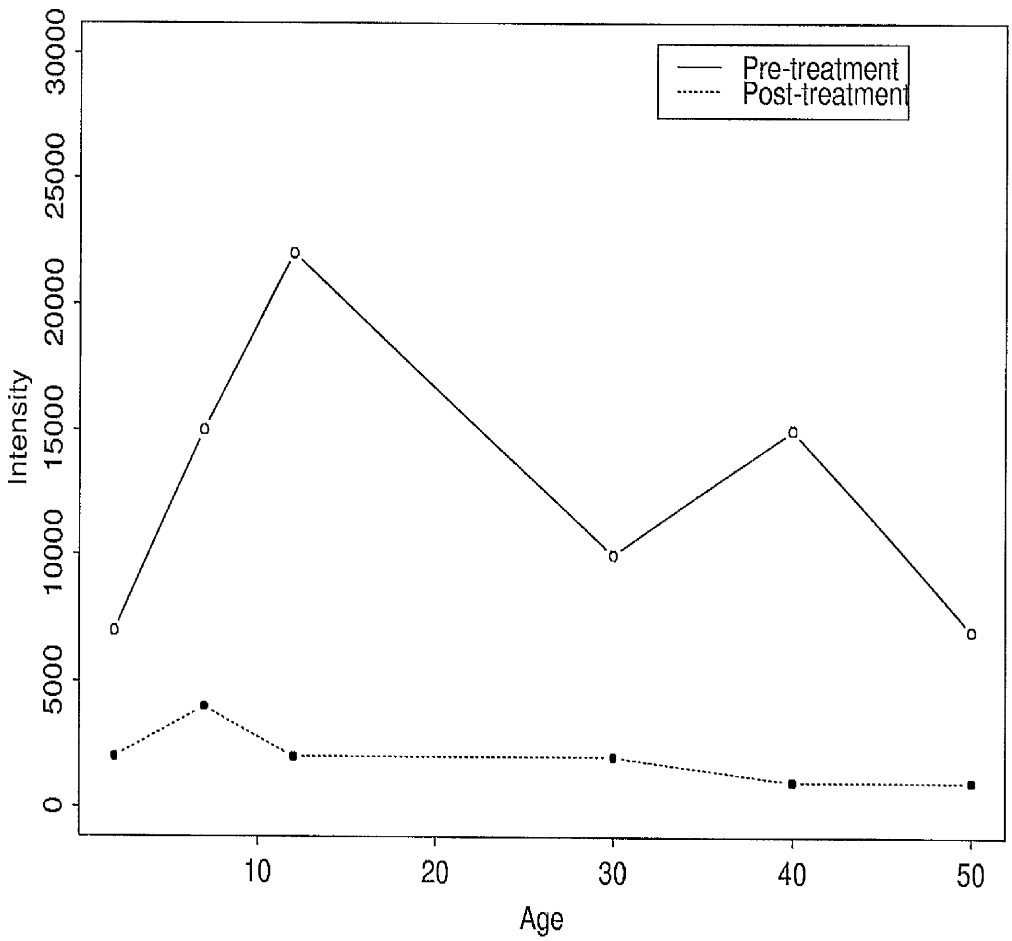


This indicates that older people do not harbor as many worms as children although they may be infected. Once again, the main reason is that the children are exposed to the contaminated soil more than the adults, thus they are expected to harbour more infected eggs. The same trend was observed in Malaysia where children with the highest prevalence of ascariasis also had the highest intensity of infection. The pattern is similar in Kenya, although the prevalence in not always highest in preschool age. (M.G. Schultz(1985)).

In order to control the transmission of the disease, three different kinds of chemotherapy treatment are provided, as has been mentioned in the previous section. The mass, the targeted and the selective treatment. A study in Nigeria compared the effects of these three treatments. It took place in four villages, Alakowe, Akeredolu, Iyanfoworogi and Iloba, which are located in Oyo State. Different types of chemotherapy treatment were provided in the first three villages, while the fourth one acted as control where no treatment was given. So, at Alakowe, selective treatment was provided in the 20% of most heavily infected individuals, at Iyanfoworogi children aged from 2 to 15 years were treated, comprising the group for targeted treatment and at Akeredolu, mass treatment was given to all residents aged more than 1 year. Two stool samples were collected, the first one in January and February 1989 before the treatment and the second one in March 1990 after the treatment. The results showed that mass and targeted chemotherapy were more effective than the selective one. In fact, the prevalence of ascariasis reduced significantly in Iyanfoworogi and Akeredolu where targeted and mass chemotherapy was provided respectively. Similar to prevalence, the intensity was reduced significantly to all three experimental villages. Even in Iloba where no treatment was given, a reduction in intensity was observed but didn't reach statistical significance.

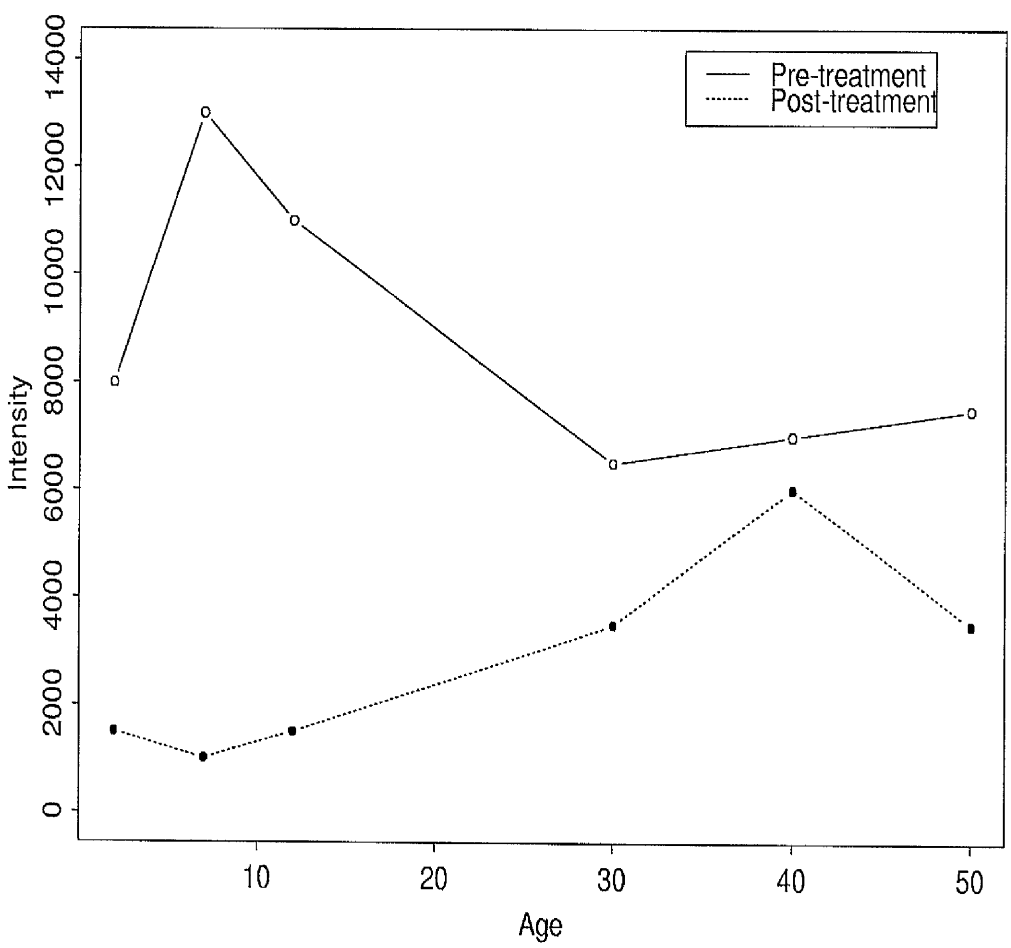
An interesting aspect was observed when the relationship between intensity and host age was examined. In Akeredolu, where mass treatment was provided, there was no significant relation between intensity and host age as we can see in figure 1.4.

Figure 1.4: Intensity of ascariasis by age in Akeredolu, Nigeria.



In fact, the intensity of ascariasis(i.e. the mean number of Ascaris worms over all people in the study) has similar pattern for all ages. However in Iyanfoworogi, where targeted chemotherapy was provided to children aged from 2 to 15 year, a totally different picture is observed (Figure 1.5)

Figure 1.5: Intensity of ascariasis by age in Iyanfoworogi, Nigeria.



As the above graph shows, there is a significant difference in intensity between young and old ages with high values for the people older than 15 years. This trend can be explained by the fact that only children from primary school were treated. However, the intensity for older ages is still lower than the pre-treatment one. The conclusion was that, whatever strategy is adopted from the community, further research which will consider the cost of each treatment in terms of man hours, transportation and material input has to be carried out. (Asaolu(1991)).

However the possibility of re-infection is inevitable. In Burma, the infection rates due to

Ascaris reached the pre-control levels in children by the sixth month after the treatment. In some communities of Malaysia, where the environmental and socio-economic factors favor the persist of ascariasis, re-infection occurs easily and rapidly. A study of 145 children from urban slums showed that the prevalence of ascariasis reached the initial levels just two months after the treatment in 21.9% of the treated children. The percentage of re-infection increased rapidly in the following months and reached the peak (80%) by the end of 5 to 6 months after the treatment.

Irrespective of prevalence, the intensity of ascariasis may be reduced after treatment. Although the infection rate after the chemotherapy reached the same levels as before, the mean burden of eggs in the contaminated soil can show a remarkable reduction after the use of anthelminthic drugs. A study in Nigeria was carried out to compare the effects of different frequencies of targeted chemotherapy with levamisole as an action for the control of Ascaris in children. Targeted chemotherapy was provided in three different communities once, twice and three times a year in children aged 5 to 15 years, a group which is at a high risk of infection. The results showed that when the drug was used at 4-monthly intervals, there was pronounced statistically significant reduction in Ascaris intensity. This reduction was potential even for untreated groups. This trend can be explained either by a real motivation of the community to help and cure the other age groups after observing the advantageous results in children. Another benefit of the targeted chemotherapy which makes it more popular is the fact that it is easy to reach children through their attendance at school, so the parents respond positively to this action. (Holland, Asaolu, Crompton, Whitehead and Coombs (1996)).

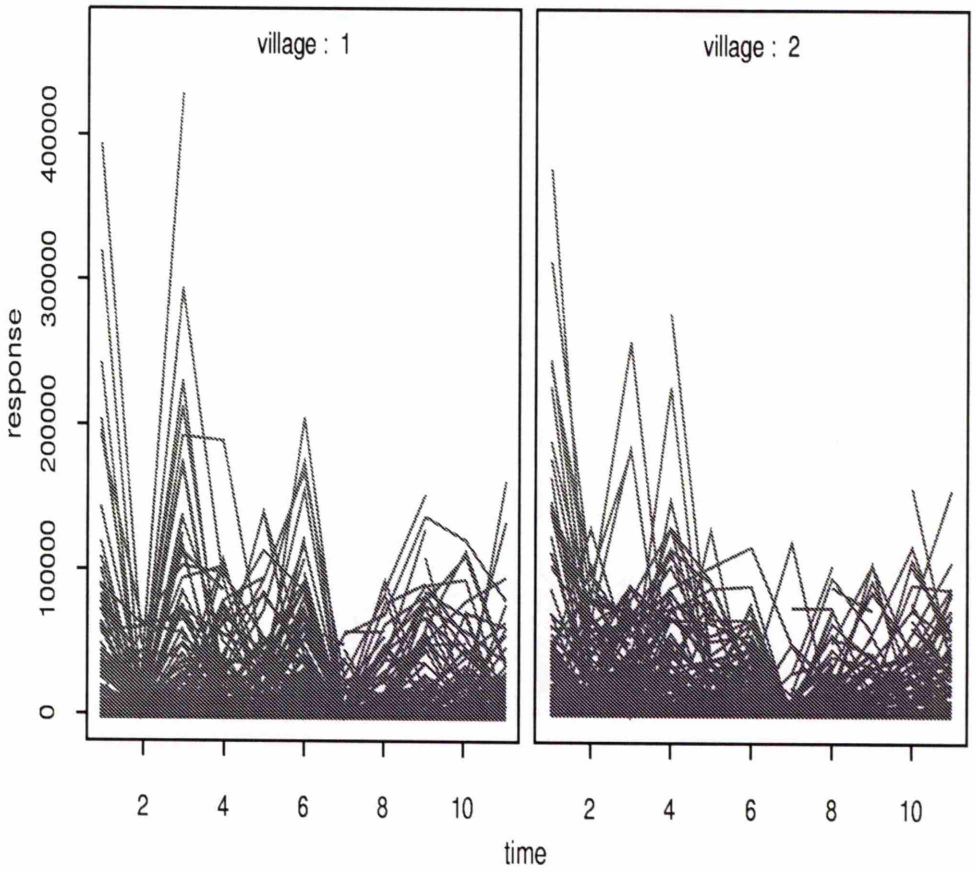
Finally, the use of multi-chemotherapy against soil-transmitted helminthiases is under discussion. However, this scheme requires more investigation and research since it is not known yet how the simultaneous dose of three drugs can effect the metabolites of the candidates after absorption into their body.

1.4 New Data

Following on from previous studies, we investigate the prevalence and intensity of ascariasis in children and adults for both sexes. In the following sections, we show how these fluctuate from younger to older ages during a period of time in which chemotherapy treatment is provided to control the infection rate of the disease. The research consists of two studies, one in Jiangxi province of China and another one in Oyo State of Nigeria where ascariasis is widely distributed. These datasets are unusual in that information is available on the same individuals from repeated measures over time. This allows longitudinal models to be fitted and analyzed.

In the Chinese study area, where the population density is 236 people per km^2, the annual temperature is between 16 and 20 ^oC and the annual rainfall ranges from 1200 to 1900mm, the second highest prevalence of ascariasis in China (71.1%) has been noticed. About 26 million cases of infection are estimated from a population of 37 millions. We focus on a cohort study of 696 people from Panchi and Laochi, two villages of Jiangxi region. 376 of them come from Panchi and 320 from Laochi with a sex ratio of about 1:1. This study started in June 1993 and finished in July 1995. It consists of two investigations, one from June'93 to June'94 and the other one from September '94 to July'95, which is the period following a mass chemotherapy treatment. We are interested in the fluctuation of the prevalence and intensity of Ascaris over the time for all age groups of both sexes. The diagnosis is made by taking samples almost every two months from the stools and measuring either the proportion of hosts found to be infected or the number of worms per person or per infected person. The following graph shows roughly an overview of the intensity of ascariasis over the time in both villages, as measured on the individuals involved.

Figure 1.6: Intensity of ascariasis by time in Panchi(1) and Laochi(2).



As we can see from the above graph, the number of eggs per infected person is quite similar for both villages and it decreases with time. This trend is expected since the treatment is applied after the sixth survey. Later on, we investigate how the intensity and prevalence are related to host age and sex over time. We will also estimate the size of the treatment effect and the time for prevalence and intensity needed to reach the pre-control levels as a result of re-infection.

The other study took place in four villages named Iyanfoworogi, Akeredolu, Alakowe and Ladin located in Oyo State of Nigeria. The average temperature in this region is between 20 and 32 ^oC, the annual rainfall ranges from 1000 to 4000mm and the vegetation is rain forest. The houses are built with mud and roofed with iron sheets while there is no organized sewage disposal system and the people get the water from ponds and streams. Unlike the study in China, we took one sample before the treatment in June/July 1991 and another one year after the treatment in July 1992. During this period, anthelminthic treatment with the drug levamisole was provided to all children attending the primary school with different frequencies for each village. So, at Iyanfoworogi the children were treated once a year in July '91, at Akeredolu twice at six-monthly intervals (in July '91 and January '92) and at Alakowe treatment was given three times at four-monthly intervals (i.e. July '91, November '91 and March '92). The village Ladin acted as a control village, in which no treatment was provided during the period of the study. Similar to the Chinese villages, we are going to examine the relation of intensity and prevalence of the disease with age and sex before and after chemotherapy. We will also estimate the size of the treatment effect, make comparisons among the different frequencies of targeted chemotherapy and investigate whether the infection rates reached the pre-treatment levels or not.

Finally, we compare the results among Chinese and Nigeria villages to assess which one of the two types of treatment, the mass or the targeted chemotherapy is the most effective for the control of soil-transmitted helminthiases such as Ascaris lumbricoides.

From a modelling perspective, an aim of the statistical work is to incorporate, where possible, an appropriate correlation structure for the measurements on a single individual across time. In some cases this is not feasible due to the high frequency of missing data. A second aim is to use smoothing techniques to explore trends in the data in a nonparametric manner.

Appropriate statistical methods are considered in Chapter 2 while covariate and treatment effects are investigated in Chapters 3 and 4 respectively. In Chapter 5, the methodology of generalized additive models is used to explore possible non-linear effects in a model comparison setting.

SECTION TWO

Repeated measures models and smoothing techniques for longitudinal data

Chapter 2

Repeated Measures models and Smoothing Techniques for Longitudinal Data

2.1 Introduction

As it is mentioned in the previous chapter, the primary aim is to investigate the prevalence and intensity of ascariasis with data obtained from measurements taken over one or two years on the same individual. In this chapter, we will describe the theoretical tools for analysis of the study, which enables us to analyze such a data set. A wide variety of well-established procedures exist for longitudinal data where the response can be modelled by a binary distribution. These are indicated briefly in the following section. For data with binary response, for example to indicate the presence or absence of Ascaris, longitudinal models have been the object of much recent research. Here we adopt one proposed by A. Azzalini (1994) for modelling repeated measures data while respecting the correlation between adjacent observations. This is described in section 2.3. In addition, we describe the construction of non-parametric regression curves. An example with data from Panchi village of Jiangxi province of China illustrates the application of the parametric model of Azzalini along with the non-parametric regression curves.

2.2 Analysis of repeated measures data with continuous responses

Repeated measures arise when measurements are taken over time or space on the same individual, i.e. a measurement could be taken before a treatment is applied, then taken again afterwards to assess the effect of the treatment and continued over some follow-up period. A commonly used method of analysis for repeated measures that involve a number of time periods is to compare the means at each time point by using either t-tests or some nonparametric equivalent. This approach could be useful if the time periods are few and the intervals between them are large. However, such tests do not give an overall answer to whether or not there is a treatment difference among the cases and do not provide a single estimate of the treatment effect.

Alternatively, we can use the method known as response feature analysis. This involves the use of summary measures, that is the responses for each individual are used to construct a single number that summarizes some features of the response profile of the individual. This method has several advantages such as: i) The appropriate choice of summary measures ensures that the analysis is focused on relevant aspects of the data, ii) It is statistically valid and iii) missing and incomplete observations can be accommodated. (Matthews (1993)). Hence, having identified a suitable summary measure we can use a t-test or a non-parametric equivalent to the single measure now available for each subject. In addition, pre-treatment measures, if available, can be used in association with the response feature approach in several ways. These involve: i) post treatment analysis, which is a simple analysis using the mean for each subject's post treatment responses on the summary measure; ii) analysis of change scores, which analyzes the differences between the measures and the individuals mean baseline measurements; iii) analysis of covariance(ANCOVA), which uses the mean of the baseline measures as a covariate. The latter could be the most powerful approach especially when the correlation between the repeated measures is not small. However, the main disadvantage is the difficulty to specify in advance an appropriate and relevant summary measure. For instance, if we are looking whether the rate of change of outcome variable differs between groups then we have to define as a summary measure the regression coefficient. On the other hand, if we are interested in whether the overall value of outcome variable is the same in different groups, then we could take as a summary measure either the overall mean for equal time intervals or the area under curve for unequal time intervals. A further problem in using the response feature approach arises when we assume that the individual summary measures in each group are identically distributed. This assumption is not always valid. For example, one can use a t-test for the difference in the treatment group means of the linear regression slopes for each measure. The t-test assumes that the slopes have identical normal distributions within groups. However, if there is a natural variation in the slopes between individuals even in the same groups, then such an assumption is not valid. The distributions of the individual slopes will have different means. Finally, ignoring that a summary measure is not a single observation but an aggregate of data from a subject may involve a loss of information that could be used to allow a more efficient analysis.

Another method of modelling repeated measures data is the analysis of variance, which involves the univariate and multivariate approach. In the former, the F-tests are valid only under some assumptions such as normality of the response variable and compound symmetry, i.e. in the covariance matrix of the repeated measures data, the variances on the main diagonal should be all equal and the covariances of each pair of data should also be equal. On the other hand, the multivariate approach does not make any assumptions about the form of the covariance matrix of the observations, but it has a low power when compound symmetry is invalid.

Otherwise, the most appropriate approach to deal with missing or incomplete repeated measures is to construct a full model for the responses, including their correlation over time, and to use the method of maximum likelihood to estimate the parameters of interest and their standard errors. The advantages of this method are as follows: i) Under the assumed model, the maximum likelihood estimates are consistent and efficient for large samples, ii) The estimation of maximum likelihood deals directly with problems of missing data because it does not require a rectangular data matrix and iii) the standard errors of parameter estimates based on the observed or expected information matrix are available and take into account that the data are incomplete. However, there are some disadvantages of this method such as: i) maximum likelihood estimation requires the specification of a full statistical model for the data and the results may be sensitive to departures from model assumptions such as normality and ii) maximum likelihood inferences are based on large sample theory and hence may be unsuitable for small data sets (Gornbein et al. (1992)).

2.3 A model for repeated measurements with binary responses

A. Azzalini (1994) proposed a model to estimate the expectation of a binary response which depends only on covariate effects . It has the form:

$$\text{logit}(\theta_t) = \mathrm{x}_t'\beta, t = 1, ..., T$$

where x_t is a vector of k time-dependent covariates, β is a k-dimensional vector of the parameters of interest and θ_t is the expected value of the response at time t. This model takes into account the correlation between adjacent observations through the parameter:

$\psi = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$, where $p_j = pr(Y_t = 1|Y_{t-1} = j)$ for $j = 0, 1$ and Y_t, Y_{t-1} are the observations at time t and t-1 respectively. The parameter ψ is known as the odds-ratio and measures the association between adjacent observations. When $\psi = 1$ the observations are independent.

The Model Of Association

We search for a parametrisation such that $\theta = E(Y_t)$ is free from the parameter that regulates the serial dependence. We distinguish two cases:

a) **Stationary case**

In this case we obtain (p_0, p_1) by solving the equations:

$$\begin{cases} \theta = E(Y_t) = \theta p_1 + (1-\theta)p_0 \quad (1) \\ \psi = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \end{cases}$$

with respect to p_0 and p_1 for given values of (θ, ψ), where θ is the probability of the response (Y=1), 1-θ is the probability of non-response (Y=0) and ψ is the odds-ratio as defined above.

b) **Non-stationary case**

In this case, which is the most commonly used in practice, we consider the generalisation of (1), which is $\theta_t = \theta_{t-1}p_1 + (1-\theta_{t-1})p_0$ and by solving the equations:

$$\begin{cases} \theta_t = \theta_{t-1}p_1 + (1-\theta_{t-1})p_0 & t = 2, ..., T \\ \psi = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} & \end{cases}$$

where $p_j = pr(Y_t = 1|Y_{t-1} = j)$ for $j = 0, 1$

we obtain:

$$p_j = \begin{cases} \theta_t, & \psi = 1 \\ \frac{\delta-1+(\psi-1)(\theta_t-\theta_{t-1})}{2(\psi-1)(1-\theta_{t-1})} + j\frac{1-\delta+(\psi-1)(\theta_t+\theta_{t-1}-2\theta_t\theta_{t-1})}{2(\psi-1)\theta_t(1-\theta_{t-1})}, & \psi \neq 1 \end{cases}$$

where $\delta^2 = 1 + (\psi-1)(\theta_t - \theta_{t-1})^2\psi - (\theta_t + \theta_{t-1})^2 + 2(\theta_t + \theta_{t-1})$ and $t = 2, ..., T$.

Hence, on taking $pr(Y_1 = 1) = \theta_1$ we generate via a nonhomogeneous Markov chain with the above probabilities a sequence Y_2,...,Y_T, so that $E(Y_t) = \theta_t$, $t = 1, ..., T$ and the odds-ratios for (Y_{t-1}, Y_t) equal to ψ.

Then the parameter $\lambda = log\psi$ shows the type of the association

a) For $\lambda = 0$ there is no association, i.e. Y_{t-1}, Y_t are independent.

b) For $\lambda \neq 0$ there is positive or negative association depending on the sign of λ.

Fitting the Model

The log-likelihood for β and λ is:

$$\mathbf{l}(\beta, \lambda) = \sum_{\mathbf{t=1}}^{\mathbf{T}} \{\mathbf{y_t}\mathrm{logit}(\mathbf{p_{y_{t-1}}}) + \log(\mathbf{1} - \mathbf{p_{y_{t-1}}})\}$$

with p_j defined as above for the 't>1' terms of summation and $p_{y_0} = \theta_1$ (Azzalini(1994), pp 768-769).The estimates for β and λ are obtained by solving the equations:

$$\begin{cases} \frac{\partial l}{\partial \beta} = 0 \\ \frac{\partial l}{\partial \lambda} = 0 \end{cases}$$

It can be shown that the ML estimator $\hat{\beta}$ is unbiased for β and has the chi-square distribution with degrees of freedom p, the dimensionality of β, i.e.

$$(\hat{\beta} - \beta)^T J^{-1} (\hat{\beta} - \beta) \sim \chi^2_p$$

where J^{-1} is the inverse Fisher Information matrix.

Then, we can estimate the standard errors and construct confidence intervals for the parameters of interest. The standard errors of the estimates are obtained by approximating the variance of the estimates given by:

$$v(\hat{\beta}, \hat{\lambda}) = \left[\sum_{i=1}^{n} \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \lambda} \end{pmatrix} \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \lambda} \end{pmatrix}' \right]^{-1}_{\beta=\hat{\beta}, \lambda=\hat{\lambda}}$$

The quantity inside the square brackets approximates the Fisher Information matrix J.

It follows that a 95% interval estimate for β is given by

$$\hat{\beta} \pm 1.96 C^{1/2}$$

where $C = J^{-1}(\hat{\beta})$ the inverse of the expected information evaluated at $\beta = \hat{\beta}$.

Missing data

Another aspect we have to look at is the problem of the missing data among the observations. Considering this problem, we assume that we have observations which are missing at random (MAR). This means that pr(Y_i is missing) does not depend on the value of Y_i. If the observations between time t and $t - m$ are missing, the log-likelihood function will be:

$$l(\beta, \lambda) = \sum_{t=1}^{T} l_t(\beta, \lambda) = \sum_{t=1}^{T} \{y_t \text{logit}(p_{t,y_{t-m}}^{(m)}) + \log(1 - p_{t,y_{t-m}}^{(m)})\}$$

where $p_{t,j}^{(m)} = pr(Y_t = 1 | Y_{t-m} = j)$, that is we replace the one step probabilities p_j with the m step probabilities $p_j^{(m)}$ and estimate the log-likelihood for the observed data. Then the first derivatives $\frac{\partial l}{\partial \beta}$ and $\frac{\partial l}{\partial \lambda}$ are obtained by computing the derivatives of $p_{t,j}^{(m)}$. Using the Chapman-Kolmogorov identity:

$$p_{t,j}^{(m)} = (1 - p_{t-1,j}^{(m-1)}) p_{t,0} + p_{t-1,j}^{(m-1)} p_{t,1}$$

we find the derivatives of $p_{t,j}^{(m)}$, which are:

$$\frac{\partial p_{t,j}^{(m)}}{\partial \omega} = -\frac{\partial p_{t-1,j}^{(m-1)}}{\partial \omega} p_{t,0} + (1 - p_{t-1,j}^{(m-1)}) \frac{\partial p_{t,0}}{\partial \omega} + \frac{\partial p_{t-1,j}^{(m-1)}}{\partial \omega} p_{t,1} + p_{t-1,j}^{(m-1)} \frac{\partial p_{t,j}}{\partial \omega}$$

where $\omega = (\theta_t, \psi)$ for $t = 1, ..., T$. In a similar manner to the complete data case, we solve the equations $\frac{\partial l}{\partial \beta} = 0$ and $\frac{\partial l}{\partial \lambda} = 0$ and we obtain the MLE for (β, λ). The disadvantage of this approach is that there are problems of convergence when a lot of data are missing.

An alternative approach is to apply the EM algorithm for ML estimation. The key idea of EM is that the missing information is not the missing observations Y_{mis} but the functions of Y_{mis} appearing in the complete data log likelihood, which is $l(\theta|Y)$ with $\theta = (\beta, \lambda)$. Specifically, the E step of EM finds the expected log likelihood of $\theta^{(t)}$, the current estimate of θ:

$$Q(\theta|\theta^{(t)}) = \int l(\theta|Y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) dY_{mis}$$

Then the M step maximizes this expected log likelihood and finds $\theta^{(t+1)}$ so that:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta_{(t)})$$

for all θ. When there is a lot of missing data, the EM algorithm converges but a lot of computation time is needed.

2.4 Non-parametric regression

Parametric regression is a widely used statistical tool because it is not only well developed but also well understood. However, there are some cases where it can not be applied because of the nonlinearity of the data. In such cases, non-parametric regression can be applied and provide a means of modelling these data. Non-parametric regression can also be used in a more exploratory manner to investigate the slopes of regression relationships without making parametric assumptions.

In non-parametric regression, the data are described by a model:

$$y_i = \mu(x_i) + \epsilon_i, (1)$$

where y_i is the response variable, x_i the covariate and ϵ_i the independent error with zero mean and variance σ^2. The function μ is assumed only to be smooth. The general problem is to estimate a mean of the response variable locally. We distinguish two cases:

a) The response variable is continuous

We define the kernel function $w(x_i - x, h)$ as a set of local weights to produce an estimate at each target value. This function is smooth and has a mode at 0 to ensure that most weight is given to the observations whose covariate values x_i lie close to x. It has the form:

$$w(x_i - x, h) = \frac{c_0}{n} d(\frac{x_i - x}{h})$$

where $d(t)$ is an even function decreasing in $|t|$, h is the bandwidth, which controls the width of the kernel function and subsequently the degree of smoothing applied to the data and c_0 is a constant chosen so that the weights sum to unity. A natural candidate for d

is the standard Gaussian density. Hence, a simple kernel approach to average the values of the response variable locally is to construct the local mean estimator

$$\mu\ (x) = \frac{\sum_{i=1}^{n} w(x_i - x; h) y_i}{\sum_{i=1}^{n} w(x_i - x; h)}$$

which was first proposed by Nadaraya(1964b) and Watson(1964). When the smoothing parameter h is the standard deviation of the normal distribution, observations with values greater than $4h$ in the x axis will contribute little to the estimate. As h increases, the resulting estimator misses some details in the curve estimate, while as h decreases, the estimator begins to track the data too closely and will end up interpolating the observed points.

An alternative approach is to construct the local linear estimator by fitting a local linear regression. We can do that by solving the least squares problem

$$min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; h)$$

and taking as the estimate at x the value of $\hat{\alpha}$, as this defines the position of the local regression line at the point x. Once more, it is the kernel weights which ensure us that most weight is given to the points close to the area of interest. Hence, the local linear estimator will take the form:

$$\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{s_2(x; h) - s_1(x; h)(x_i - x)\} w(x_i - x; h) y_i}{s_2(x; h) s_0(x; h) - s_1(x; h)^2}$$

where $s_r(x; h) = \frac{\sum_{i=1}^{n} (x_i - x)^r w(x_i - x; h)}{n}$.

At this point, we can notice that the local mean estimator can be derived in a similar way, by removing the $\beta(x_i - x)$ term from the equation of the least squares problem. The use of local linear estimator is appealing because it can be viewed as a relaxation of the usual linear regression model. As the smoothing parameter h becomes very large, the weights attached to each observation by the kernel function become very close and the curve estimate approaches the fitted linear regression line.

In general, $\hat{\mu}(x)$ is a biased estimator for $\mu(x)$, but the bias will be small if the smoothing parameter h is small so that the weights decay quickly to zero as we are moving away from the location of interest. Also the variance of $\hat{\mu}(x)$ increases as h decreases. The need to balance bias against variance is characteristic of non-parametric smoothing problems, which explains why the choice of the smoothing parameter h is of primary interest.

One way to choose h is automatically from the data by using the mean squared error criterion. In this case, we choose h to minimize the quantity

$$S(h) = \sum_{i=1}^{n} \{y_i - \hat{\mu}_{-i}(x_i)\}^2$$

The reason for minimizing the above quantity is that it is estimating the mean square error of $\hat{\mu}(x)$ for $\mu(x)$ averaged across the design points x_i up to an additive constant which does not depend on h.

In practice, a sensible choice of h is one which tends to zero as the number of subjects increases, in which case the bias becomes negligible and we can estimate the variance of $\hat{\mu}(x)$ and hence the standard errors for the estimates $\hat{\mu}(x)$.

b) The response variable is binary

In this case, the use of a local linear estimator, provided by fitting a local linear regression, has some undesirable features such as:

i) The motivation for $\hat{\mu}(x)$, which was provided by (1), cannot held in this case, if ϵ is to be an error term in the traditional case.

ii) The fitted curve is not guaranteed to lie in the interval (0,1).

iii) It is not appropriate for satisfactory evaluation of standard errors, when these are required.

To overcome these difficulties, we adopt linear logistic regression as the local model and apply weights to the log-likelihood of the form:

$$l_{(h,x)}(\alpha,\beta) = \sum_{i=1}^{n} l_i(\alpha,\beta) w(x_i - x; h)$$

where $l_i(\alpha,\beta) = y_i log(\frac{p_i}{1-p_i}) + log(1-p_i)$, p_i = probability of 1 at x_i and $log(\frac{p_i}{1-p_i}) = \alpha + \beta x_i, i = 1, 2, ..., n$.

We then maximize the $l_{(h,x)}(\alpha,\beta)$ with respect to (α,β) and obtain local estimates $(\hat{\alpha},\hat{\beta})$ and fitted value:

$$\hat{\mu}_i(x) = \frac{exp(\hat{\alpha} + \hat{\beta}x)}{1 + exp(\hat{\alpha} + \hat{\beta}x)}$$

This procedure is repeated for a sequence of x values and a set of pairs $(x, \hat{\mu}(x))$ is constructed and plotted (Bowman & Azzalini, Applied Smoothing Techniques for data analysis, pp 117-23).

Again, the same principles apply as in the case of data with a continuous response. As the smoothing parameter h increases the bias increases and the variance decreases.

Throughout our research, the standard method to use will be the local linear approach to the data with a continuous response and the smooth logistic regression to the data with a binary response. For both cases, we use a constant smoothing parameter h, which has been chosen subjectively.

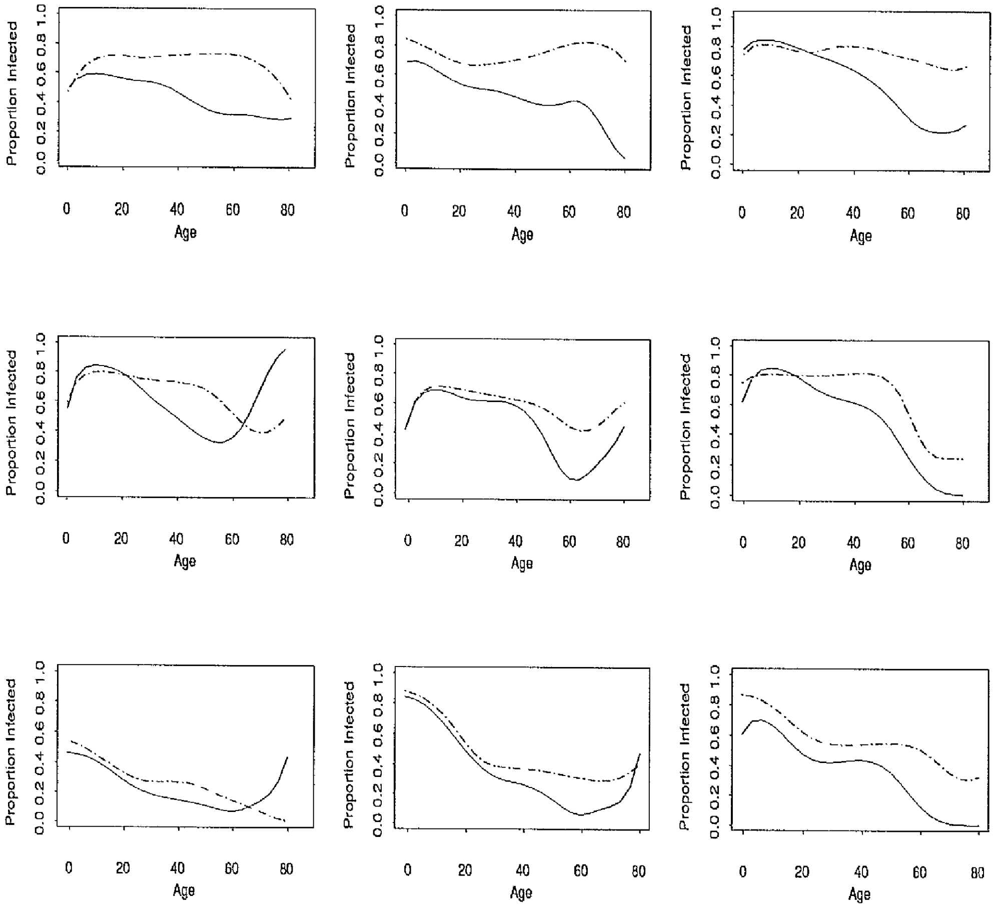
2.5 A real data example

Here we illustrate the way in which we can use the parametric logistic regression technique along with the non-parametric smoothing curves described above. The data are obtained from a cohort study of 343 people from Panchi village in Jianxgi Province of China, started in June '93 and finished in January '95. Each patient was examined over several occasions and at each occasion the presence or absence of infection was noted.

Figure 2.1 shows the proportion infected by Ascaris before (the first six graphs) and after the treatment (the last three graphs), as a function of age and sex. The slopes indicated in these graphs are simple curve-fitting moment estimates with a smooth function for age and a common value of 10 for the bandwidth h. Note that in our case, the aim is descriptive, so the value of h has been chosen subjectively.

The beneficial effect of treatment for both males and females not only in children but also in adults, is apparent from these graphs.

Figure 2.1: Prevalence of ascariasis in Panchi village of China for males(Continuous line)and females(Dashed line). Pre-treatment period: June'93 - June'94(first six graphs), Post-treatment period: July'94 - January '95(last three graphs)



Fitted Models for Prevalence

In order to interpret the non-parametric curves, we fit a parametric logistic regression model for comparison. For the case of a binary response, we consider the model proposed by Azzalini as it has been described in section 2.3, which is a logistic regression model with the addition of serial dependence of Markovian form. In such a model Age, Sex and Time are being fitted as factors with 3 (i.e. children 0 to 5 years, children 5 to 15 years and adults), 2(i.e. Males and Females) and 9 levels respectively. We then define $k-1, (k=3 \text{ or } 9)$, "dummy" variables(or "design" variables or "contrasts") as follows:

Level i	$t_{21} =$	$t_{31} =$	**...**	$t_{k1} =$
1	0	0	...	0
2	1	0	...	0
3	0	1	...	0
...	....	...	...	.
k	0	0	...	1

Thus, we replace the 3 levels of age by 2 simple binary explanatory variables, say t_{21} and t_{31} respectively. In this case, the logistic regression parameters b_{21} and b_{31}(i.e. the coefficients of the two "dummy" explanatory variables t_{21} and t_{31}) are simply the differences in log odds between level 2 and level 1 and then level 3 and level 1 of the factor respectively. Similarly, we replace the 9 levels of Time by 8 simple binary explanatory variables. The resulting coefficients, $c_{21}, c_{31}, ..., c_{91}$ will be the difference in Log Odds between level 2 and level 1, level 3 and level 1,..., and level 9 and level 1 respectively. Log Odds between any other two levels can be obtained from the appropriate differences, e.g. $c_{76} = c_{71} - c_{61}$ is the log odds of response in level 7 to that in level 6. Finally, we re-code sex as a binary variable, which takes the value 0 for level 1(Males) and 1 for level 2(Females). Then we fit a full model with the log odds of the response as depending linearly on the explanatory variables.

In the following table, several models are displayed with the main effects and their interactions.

Table 2.1: Fitted models

Model	**-2Log-likelihood**	**No. of parameters**
Age+Sex	2383.2502	4
Time+Age+Sex	2228.2334	12
Time+Age+Sex+Age*Sex	2224.4536	14
Time+Age+Sex+Time*Age	2174.9112	28
Time+Age+Sex+Time*Sex	2224.755	20

We choose the model, which describes the data in a reasonable way, by using a combination of forward and backward methods based at any step on adding or deleting terms as appropriate. Whether a variable should be included in the model for the next step will be based on a simple likelihood ratio test of a model (A) without the particular variable within model (B) which includes that particular variable, i.e. rejecting model (A) in favor of model (B) if only if

$$D_A - D_B > \chi^2(p_B - p_A; 0.95)$$

, where D_A and D_B are the deviances (i.e. -2maxloglikelihood) and p_A, p_B are the number of parameters of models A and B respectively.

So, we end up with the model:

log-odds of the disease $= a + Time + Age + Sex + Time * Age$

with main effects of time, age and sex as well as an interaction effect between time and age.

Fitted Model Parameters

The table below shows the estimates of the parameters in the model which are of interest along with their standard errors, the odds-ratios and the corresponding 95% confidence intervals. At this point, we should mention that although the term of age is in a continuous scale, as it is clearly indicated from figure 2.1, we consider it as a factor with three levels and use one of them as a baseline(i.e. children aged 0 to 5 years). In this way, the resulting coefficients b_{21} and b_{31} will be the differences in log odds between level 2(children aged 5 to 15 years) and level 1(children 0 to 5 years) and between level 3(adults) and level 1(children 0 to 5 years) respectively. Also, it has to be noticed that the term of time is a factor with nine levels, as it is indicated from the graphs which show different slopes at each time point. However, in the following table, we specify only the time period which is the significant one and that is the period just immediately after treatment. The resulting coefficient for time will be c_{76}, which is the difference in log odds between level 7(post-treatment) and level 6(pre-treatment). Similarly for the interaction effects, we present only the significant ones corresponding to that period with coefficients f_{21} for the difference in log-odds between levels 2(children aged 5 to 15 years) and 1(children aged 0 to 5 years) and f_{31} for the difference in log-odds between levels 3(adults) and 1(children aged 0 to 5 years). The coefficient d_{21} represents the difference in log-odds between levels 2(Females) and 1(Males).

Table 2.2: Estimated Parameters

Variable	**Term**	**Coeff.**	**Std. error**	**t value**	**Odds ratio**	**95% CI for variable**	**95% CI for odds-ratio**
Constant	a	0.707	0.078				
Time	c_{76}	-0.221	0.025	-8.580	0.80	(-0.27,-0.17)	(0.76,0.84)
Age(1)	b_{21}	0.394	0.109	3.587	1.48	(0.18,0.60)	(1.20,1.82)
Age(2)	b_{31}	-0.353	0.046	-7.654	0.70	(-0.44,-0.26)	(0.64,0.77)
Sex	d_{21}	0.278	0.064	4.362	1.32	(0.15,0.41)	(1.16,1.51)
Time*Age(1)	f_{21}	-0.114	0.038	-2.964	0.89	(-0.18,-0.04)	(0.84,0.96)
Time*Age(2)	f_{31}	-0.055	0.016	-3.434	0.95	(-0.09,-0.03)	(0.91,0.97)

From the fitted model we confirm the significant reductions in prevalence of the disease with age and time, due to the effect of the treatment. Despite the fact that young children are more exposed to the disease, the proportion of those who are infected just immediately after the treatment is decreasing significantly with an odds-ratio of 0.89 even lower than that in adults. Although from the graph, there does not seem to be a very big difference between males and females, the fitted model suggests that females do differ from males with an increased risk of 1.32, which lies in a 95% confidence interval: (1.16, 1.51).

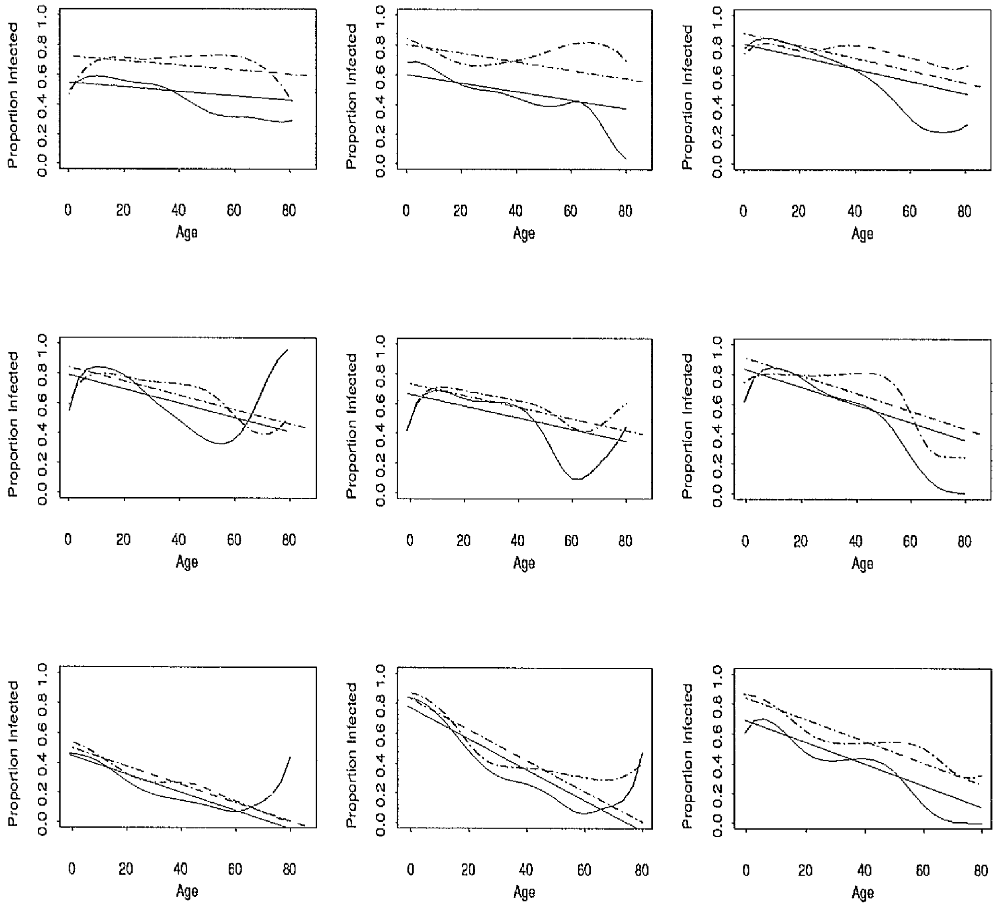
Another way to handle this dataset is to consider age as a linear effect. In this case, the estimated coefficients for age and sex as predictors for the prevalence of ascariasis in the 7th survey(post-treatment) will be:

Table 2.3: Estimated Parameters

Variable	**Coefficient**	**Std. Error**	**p value**
Constant	0.1278	0.2724	0.6388
Age	-0.0348	0.0088	0.0001
Sex	-0.2137	0.3150	0.4975

As we can see from the above table, the effect of age is significant but there is no significant sex effect for the prevalence of the disease in that period. In the following graph, we display graphically the fit of a logistic model in which age is treated as a linear effect along with non-parametric regression curves for each period of time.

Figure 2.2: Parametric and non-parametric fit for the prevalence of ascariasis in Panchi.Continuous line: Males, Dashed line: Females



In the above figure, it can be shown that the parametric model does describe the data in a reasonable accurate way since its fitted lines are quite close to the smooth curves obtained by using non-parametric regression. In the following chapters, we will try to fit analogous models to the rest of our data. Specifically, in chapter 3, we will investigate the effects of the risk factors as linear predictors of the response, while in chapter 5, we will assess the accuracy of these linear models by plotting them together with an additive model. In this way, we can explore the most effective scale on which age can be assessed.

SECTION THREE

Modelling the covariate effects

Chapter 3

Modelling the covariate effects

3.1 Introduction

In this chapter the main interest is to investigate the way in which the risk factors such as time, age and sex affect the prevalence and the intensity of ascariasis.

That is, we want to examine if the proportion of cases infected by Ascaris or the number of Ascaris worms per person or per infected person differs with age, as well as if there is any significant difference between males and females along with the effect of the treatment.

3.2 Statistical methods

As we deal with longitudinal data which means repeated measures over the time, the most appropriate way to find an association among the risk factors is to fit a parametric model, which takes into account the correlation between successive observations. Such a model has been described in chapter 2 along with an illustration in a real data set from Panchi. However, when we took a new sample from both Panchi and Laochi village, as well as one from Nigeria, we found that this model does not work, since it can not handle data which are missing in more than the 1/3 of the total number of observations. Moreover, the missingness mechanism is not MCAR (missing completely at random), for we can not delete the missing data and fit the model only to the complete data set. That would lead to invalid inferences because of the loss of information. For this reason, we will fit a generalized linear model which can handle missing observations. Although such a model does not take into account the correlation over time, its methods of inference are based on well-founded theory under the assumption of independent errors. Then, we will compare it with the one we fit in the first sample of Panchi, where there is sufficient data to allow a correlated error model to be fitted.

Once the model which describes the data properly has been selected by using the log-likelihood ratio test, it will be displayed graphically along with the non-parametric regression curves.

3.3 Results

Model for prevalence

In this model, the probability of the response variable is expressed linearly as a combination of the explanatory variables and interactions among them of the form:

$$logit(\pi(x)) = log(\frac{\pi(x)}{1-\pi(x)}) = x^T\beta$$

where $\pi(x) = pr(Y = 1|x)$. The main reason for its popularity is that the logit model ensures that the proportions π lie in the interval (0,1) without any constraints of the linear predictor $x^T\beta$. This is clear from the inverse relationship:

$$\pi(x) = \frac{exp(x^T\beta)}{1+exp(x^T\beta)}$$

The above equations allows us to draw some conclusions from the parameters based on such a model. Suppose we have two explanatory variables x_1 and x_2, so that the above equation takes the form:

$$\pi(x) = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

Then, assuming that x_1 and x_2 are functionally unrelated, a unit change in x_2 has the result of increasing the odds of the response variable multiplicatively by a factor $exp(\beta_2)$. Of course, in order to achieve this, it is important for x_1 to be held fixed. On the other hand, the interpretation of the parameters on the probability scale is more complicated because the effect on π of a unit change in x_2 depends on the values of x_1 and x_2. This is clearly shown from the derivative of $\pi(x)$ with respect to x_2:

$$\frac{\partial\pi}{\partial x_2} = \pi(1 - \pi)\beta_2$$

So, a small change in x_2 has a larger effect, as measured on the probability scale, if π is near 0.5 than if π is near 0 or 1. Therefore, the most appropriate way to draw conclusions is by plotting the proportion of individuals infected versus the effect of age and sex over time. In this way, a change in the explanatory variables will have a corresponding effect on the probability of the response.

Parameter estimation for prevalence

To estimate the parameters β, we use the method of maximum likelihood. The log-likelihood function has the form:

$$l(y,\beta) = \sum_{i=1}^{n}\{y_i log(\frac{\pi_i}{1-\pi_i}) + (1-y_i)log(1-\pi_i)\}$$

where $\pi_i = P(x_i) = exp(\beta x_i)/(1 + exp(\beta x_i))$ and x_i are the explanatory variables, i=1,2,...,n. On taking the derivative of l with respect to β and setting equal to zero, we get:

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} x_i(y_i - \pi_i) = 0$$

which are non linear in the parameters β and consequently one has to use an iterative method such as the Newton-Raphson procedure. That is, we estimate initial values for $\hat{\beta}$, which give estimated probabilities π_i. Then we construct the linearized response:

$$z_i = \hat{\beta}^T x_i + \frac{y_i - \pi_i}{\pi_i(1-\pi_i)}$$

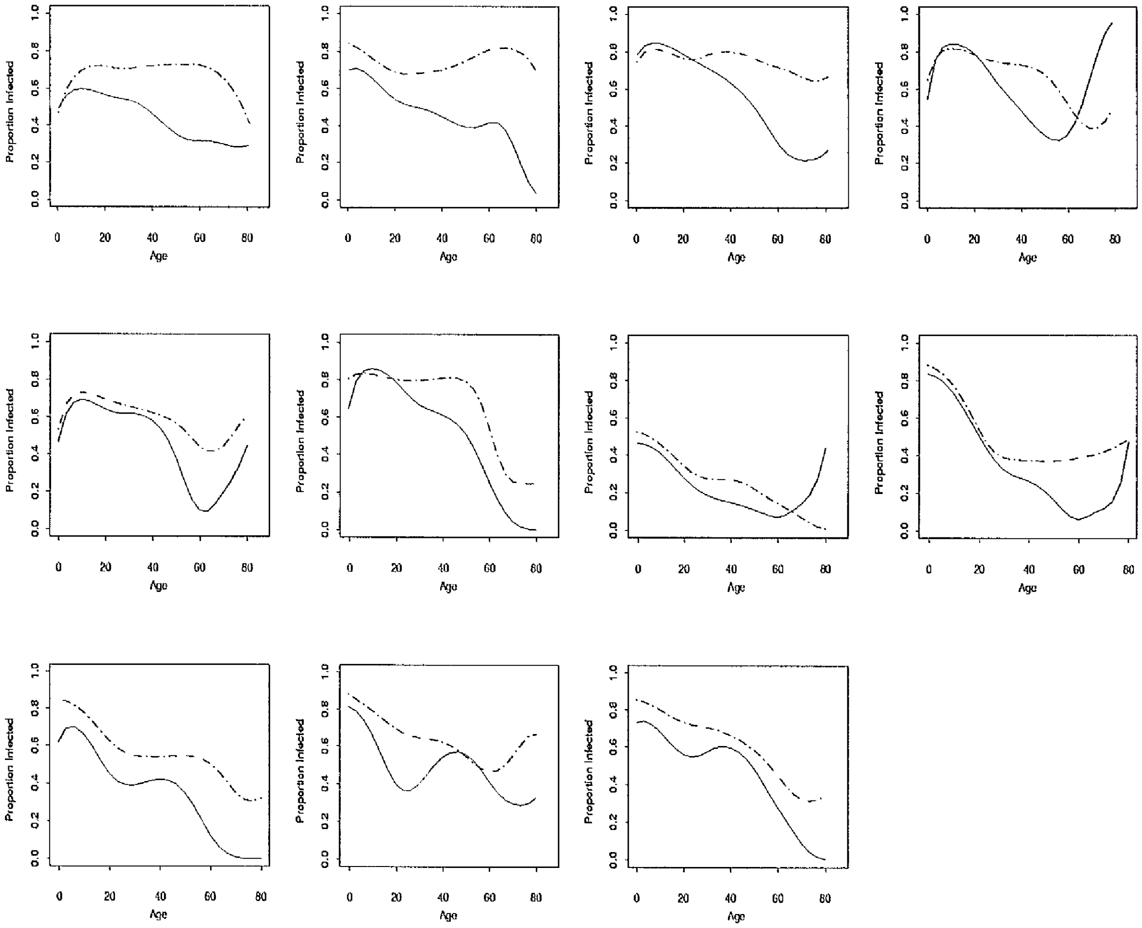
and weights $w_i = \pi_i(1-\pi_i)$. So, a new $\hat{\beta}$ is obtained by weighted linear regression of z_i onto the x_i with weights w_i. We repeat this procedure until $\hat{\beta}$ converges.

Fitting the model

Making use of the model described above, we will carry out an analysis for the prevalence of ascariasis in Jiangxi Province of China and Oyo State of Nigeria and then we will try to analyze the intensity in these countries.

The following graphs use non parametric regression curves to show how the proportion infected by Ascaris in Panchi village fluctuates with the age of both sexes from June '93

Figure 3.1: Prevalence of ascariasis in Panchi, Continues line: Males, dashed line: Females



to July '95. In this point it has to be mentioned that although figure 3.1 is identical to figure 2.1 indicated in chapter 2, there is a difference in time for the post-treatment period, in which 2 more levels have been added. Also, we have subjectively chosen the value of 10 for the bandwidth h.

As we can roughly see from the above graph, the disease is widely distributed in children but less in adults, while there seems to be a slightly higher infection for females. However in the 7th survey which is the period just immediately after the treatment, this proportion drops significantly in males and females for all the age groups. After this period, the prevalence increases again, especially in children as a result of re-infection.

A formal analysis involves the application of several models to the data. Similar to the parametric models fitted in chapter 2, we re-code sex as a binary variable, which takes the value 0 for level 1(Males) and 1 for level 2(Females. We also define $k-1, k \geq 3$ "dummy" variables for age and time. The only difference is that now time consists of 11 levels. The logistic regression parameters are also defined in a similar way to the ones in the previous chapter. We then use analysis of deviance to carry out a both ways(forward and backward) stepwise variable selection. Table 3.1 shows the different models which have been fitted along with the log-likelihood and the number of parameters.

Table 3.1: Fitted models to data from Panchi

Model	**-2Log-likelihood**	**No. of parameters**
Time+Age+Sex	2929.736	14
Time+Age+Sex+Age*Sex	2923.871	16
Time+Age+Sex+Time*Age	2860.388	34
Time+Age+Sex+Time*Sex	2923.819	24
Age+Sex	3056.967	4

By using the log-likelihood ratio test, we end up to the model:

log-odds of the disease= $Intercept + Time + Age + Sex + Time * Age$, with main effects of time, age and sex as well as an interaction effect between time and age.

The estimated coefficients along with the standard errors, the 95% CI's and the odds-ratios of each variable are shown table 3.2. As it has been mentioned in Chapter 2, time is a factor with 11 levels(6 out of them correspond to the period before treatment and the other 5 to the following up period). Once more, it may be a reasonable approximation to regard age as a factor with 3 levels. In the following table, we present only the significant time and the interaction between time and age effect corresponding to the period just immediately after treatment. The coefficients b_{21} and b_{31} represent the difference in log-odds ratios between levels 2(children aged 5 to 15 years) and 1(children aged 0 to 5 years) and levels 3(adults) and 1 respectively, c_{76} is the difference in log-odds between post and pre-treatment level and f_{21}, f_{31} the difference in log-odds for the post-treatment period between levels 2(children aged 5 to 15 years) and 1(children aged 0 to 5 years) and between levels 3(adults) and 1 respectively. The coefficient d_{21} for Sex represents the difference in log-odds between levels 2(Females) and 1(Males).

Table 3.2: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	Odds ratio	95% CI for variable	95% CI for odds-ratio
Constant	a	0.806	0.057				
Time	c_{76}	-0.220	0.025	-8.75	0.80	(-0.27,-0.17)	(0.76,0.84)
Age(1)	b_{21}	0.352	0.081	4.33	1.42	(0.19,0.51)	(1.21,1.67)
Age(2)	b_{31}	-0.397	0.033	-11.86	0.67	(-0.45,-0.33)	(0.64,0.72)
Sex	d_{21}	0.307	0.045	6.75	1.36	(0.22,0.40)	(1.25,1.49)
Time*Age(1)	f_{21}	-0.126	0.039	-3.23	0.88	(-0.21,-0.05)	(0.81,0.95)
Time*Age(2)	f_{31}	-0.053	0.016	-3.29	0.95	(-0.08,-0.02)	(0.92,0.98)

The above table shows the significant effects of age, sex and time as the effect of the treatment, in the prevalence of ascariasis. Despite the fact that young children are more exposed to the disease than adults with an odds ratio of 1.42, the effect of the treatment decreases this ratio to 0.88 just immediately after the treatment. There was also a significant reduction for adults at that period. However, females are still 1.36 times more likely to be infected than males.

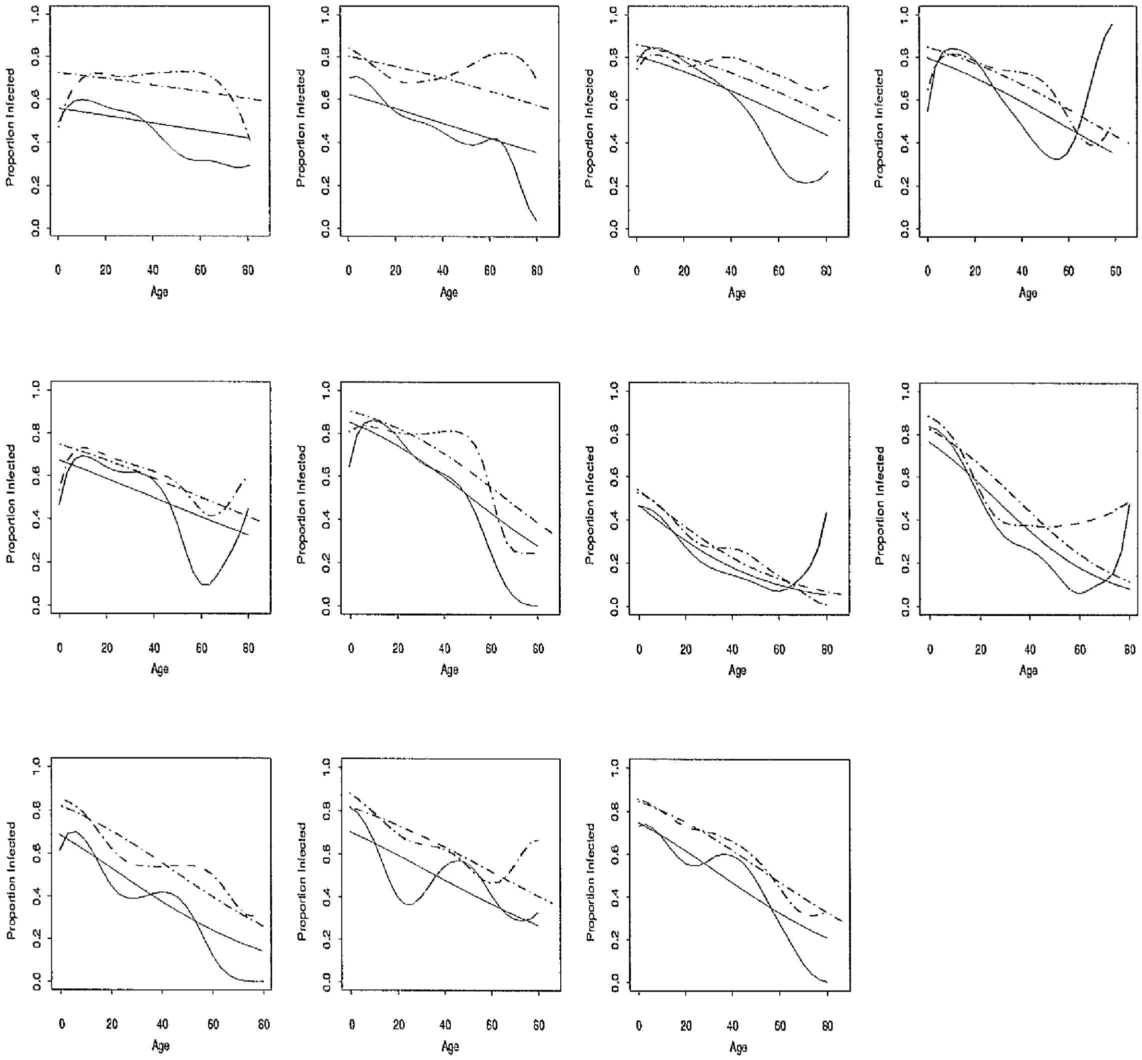
In the graph 3.2, we present the model in which age is treated as a linear effect along with the non-parametric regression curves. Similar to the first dataset from Panchi, the estimated parameters for this model and for the period just immediately after treatment are shown in the following table.

Table 3.3: Estimated Parameters for Panchi

Variable	**Coefficient**	**Std. Error**	**p value**
Constant	0.1439	0.2625	0.58
Age	-0.0343	0.0087	0.001
Sex	-0.2781	0.3092	0.368

There is a significant effect for age but the effect of sex is not a significant predictor for the presence or absence of the disease. From figure 3.2, we can clearly see that this model describes the data accurately since it is close to the non-parametric one.

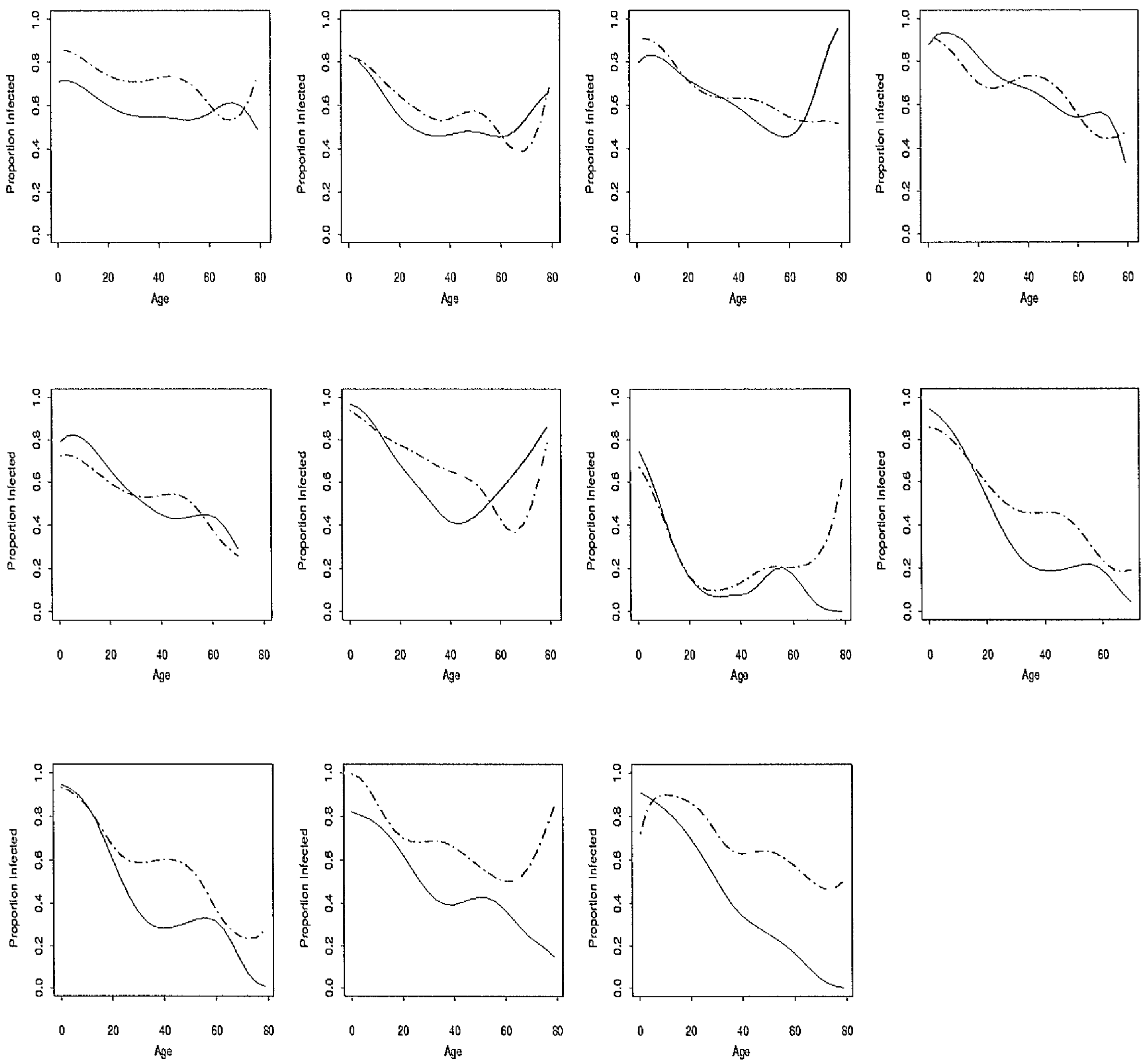
Figure 3.2: Parametric and non-parametric fit for the prevalence of ascariasis in Panchi



Comparing this model with the one we fit in the first data sample from Panchi, we see that it gives similar results although it does not take into account the correlation between successive observations. This encourages us to fit a similar model to the remaining set of data from the other villages.

As for the second village of Jiangxi province named Laochi, the prevalence of ascariasis can be shown in the following graph. It is apparent from the graph 3.3 that there is a similar pattern to Panchi. The proportion infected is higher in children but lower in adults and it is decreasing in the period following a mass chemotherapy treatment in all ages. Again, we can notice a higher trend for females.

Figure 3.3: Smooth curves of Prevalence of ascariasis to age in Laochi. The smoothing parameter is equal to 10



The next step is to carry out a formal analysis following the same procedure with Panchi village(i.e. age is a factor with 3 levels, sex is a factor with 2 levels and time is a factor with 11 levels). The different fitted models along with the log-likelihoods and the number of their parameters, are shown in table 3.4.

Table 3.4: Fitted models to data from Laochi

Model	**-2Log-likelihood**	**No. of parameters**
Age+Sex	2738.383	4
Time+Age+Sex	2604.993	14
Time+Age+Sex+Age*Sex	2600.699	16
Time+Age+Sex+Time*Sex	2590.805	24
Time+Age+Sex+Time*Age	2565.908	34

The model which describes the data properly is similar to Panchi:

log-odds of the disease= $Intercept + Time + Age + Sex + Time * Age$,

with estimated coefficients and odds-ratios as well as the corresponding 95% CI's, as they are shown in table 3.5. The resulting coefficients are displayed in the same manner as the ones from Panchi village.

Table 3.5: Fitted model parameters

Variable	**Term**	**Coeff.**	**Std. error**	**t value**	**Odds ratio**	**95% CI for variable**	**95% CI for odds-ratio**
Constant	a	1.054	0.066				
Time	c_{76}	-0.252	0.029	-8.458	0.78	(-0.31,-0.19)	(0.73,0.83)
Age(1)	b_{21}	0.011	0.094	1.07	1.01	(-0.17,0.19)	(0.84,1.21)
Age(2)	b_{31}	-0.59	0.037	-15.78	0.55	(-0.66,-0.52)	(0.52,0.59)
Sex	d_{21}	0.188	0.047	3.934	1.72	(0.10,0.28)	(1.11,1.32)
Time*Age(1)	f_{21}	-0.103	0.038	-2.687	0.90	(-0.17,-0.03)	(0.84,0.97)
Time*Age(2)	f_{31}	-0.035	0.019	-1.786	0.97	(-0.08,0.01)	(0.92,1.01)

We can see from the above table the significant effects of time as a treatment effect, age, sex as well as the interaction effect between age and time. There is a significant reduction in prevalence for children but not for adults since this group is less exposed to the disease with an odds-ratio of 0.55. It can also be noticed that females are at higher risk than males with an odds-ratio of 1.72.

Graph 3.4 shows graphically the fit of the model in which age is considered as a linear effect. Similar to Panchi, the estimated parameters for age and sex for the period just immediately after treatment are shown in the following table.

Table 3.6: Estimated Parameters for Laochi

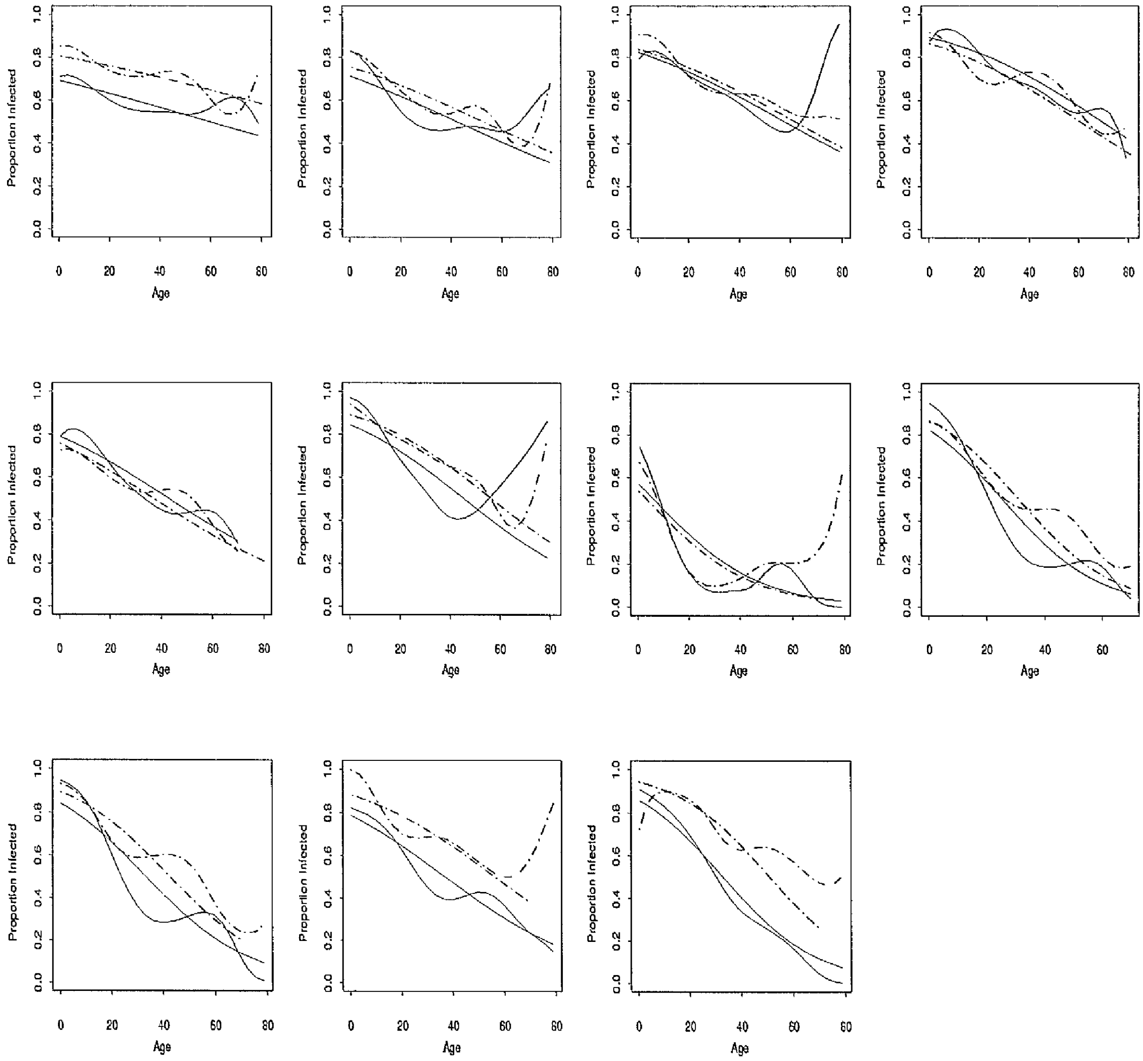
Variable	Coefficient	Std. Error	p value
Constant	0.164	0.328	0.616
Age	-0.048	0.011	0.00
Sex	0.142	0.35	0.68

Again, age is a significant predictor for prevalence in Laochi but sex is not.

It is apparent that the parametric model, in which age is treated as a linear effect does describe the data in a reasonably accurate way. Comparing the effect of the treatment in the two villages of China, we can say without doubt that it is similar for both of them. Although young children and adults seem to be more infected by Ascaris in Panchi than in Laochi for the period before the treatment, there is a slightly higher reduction in prevalence for the former just after treatment. However, females in Laochi are in higher risk than those in Panchi.

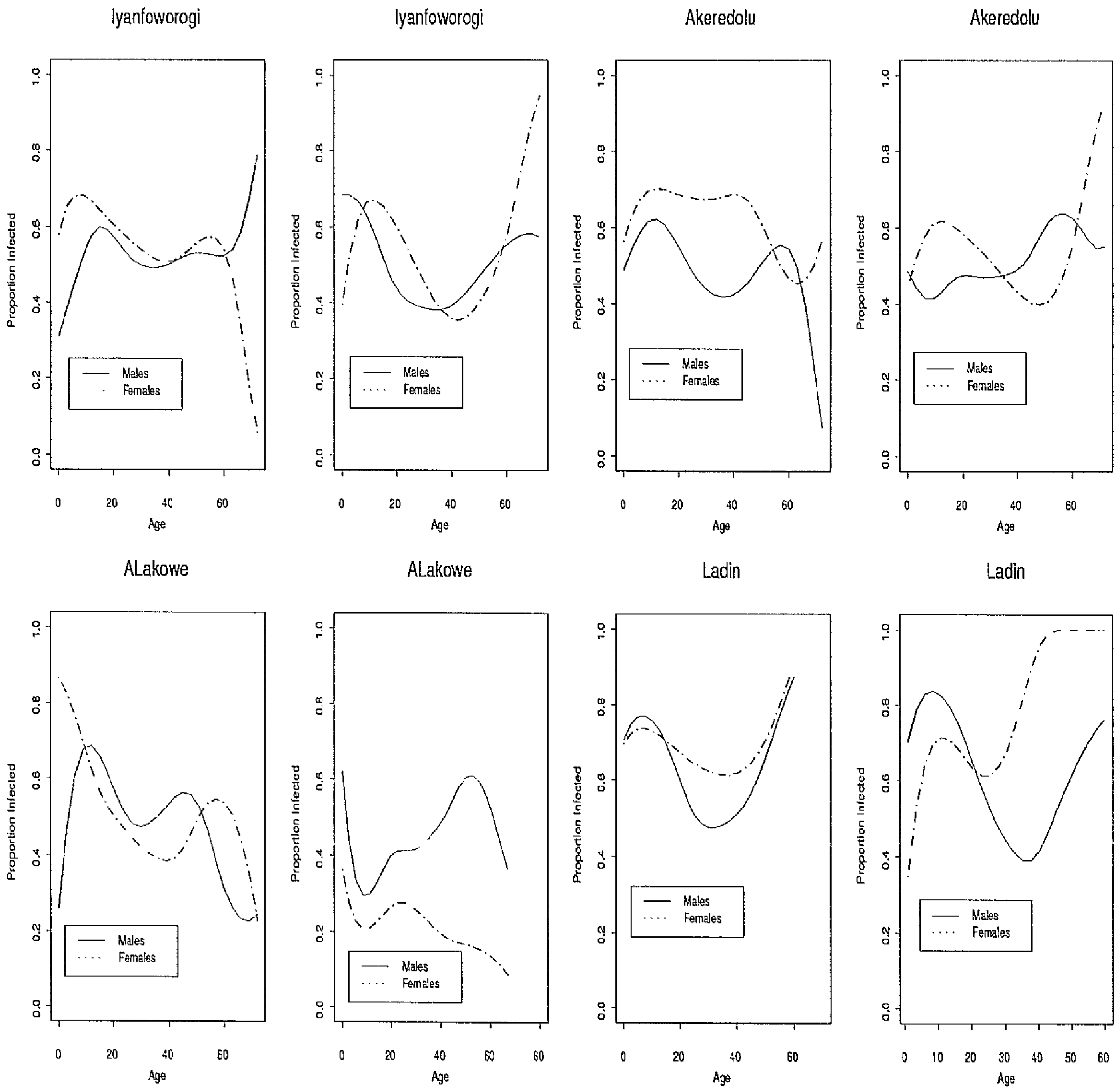
We now move on to investigate the prevalence of ascariasis in the four villages of Nigeria. As has been mentioned in the first chapter, targeted chemotherapy with the drug levamisole was provided only to the group aged between 5 to 15 years with different frequencies in each village. In Alakowe, the drug was provided three times a year at four-monthly intervals, in Akeredolu twice a year, in Iyanfoworogi once a year, while the

Figure 3.4: Parametric and non-parametric fit for the prevalence of ascariasis in Laochi



last one Ladin acted as a control village where no treatment was given. Graph 3.5 shows the proportion infected by the disease in each village before(left panel) and one year after the treatment(right panel). For each plot, the value of the smoothing parameter is set subjectively to 10.

Figure 3.5: Smooth curves of Prevalence of ascariasis to age in four villages of Nigeria. For each village left graph corresponds to the period before treatment and the right one to the follow up period



Following exactly the same procedure as before(i.e. considering age as a factor with 3 levels, sex as a factor with 2 levels and time as a factor with 2 levels), we will fit several models for each village focusing only in the group where the treatment was provided and select the one which describes the data properly. So, starting with Iyanfoworogi, the fitted models are shown in table 3.7.

Table 3.7: Fitted models to data from Iyanfoworogi

Model	**-2Log-likelihood**	**No. of parameters**
Age	537.5822	3
Age+Sex	535.8437	4
Age+Time	537.5446	4
Time+Age+Sex	535.802	5

Using the log-likelihood ratio test, we conclude to the model:

log-odds= $Intercept + Age$

with estimated coefficients, odds-ratios and the corresponding 95% CI's as shown in table 3.8. Again, the coefficients b_{21} and b_{31} represent the difference in log-odds between levels 2(children aged 5 to 15 years) and 1(children 0 to 5 years) and between levels 3(adults) and 1 respectively.

Table 3.8: Fitted model parameters

Variable	**Term**	**Coeff.**	**Std. error**	**t value**	**Odds ratio**	**95% CI for variable**	**95% CI for odds-ratio**
Constant	a	0.356	0.137				
Age(1)	b_{21}	0.167	0.141	1.186	1.18	(-0.11,0.45)	(0.90,1.57)
Age(2)	b_{31}	-0.144	0.071	-2.011	0.87	(-0.28,-0.004)	(0.75,0.99)

As is shown in the above table, although targeted chemotherapy treatment was provided in young children, there is no significant reduction in prevalence for this age group. In the contrary, there is an increase with an odds-ratio of 1.18, which however does not reach significant levels. That means that there is no evidence that targeted chemotherapy has an effect when it was provided once a year.

Pursuing our analysis, we move on to the prevalence in Akeredolu where the treatment was provided at six-monthly intervals. The fitted models are as shown in table 3.9.

Table 3.9: Fitted models to data from Akeredolu

Model	-2Log-likelihood	No. of parameters
Sex+Time	681.9715	3
Age+Sex	688.019	4
Time+Sex+Age	681.6697	5
Time+Age+Sex+Age*Sex	679.4082	7
Time+Age+Sex+Time*Age	679.2385	7
Time+Age+Sex+Time*Sex	680.6442	6

The model, which describes the data accurately is: log-odds= $Intercept + Time + Sex$ with the estimated coefficients, odds-ratios and the corresponding 95% CI's as they are shown in table 3.10. The coefficient c_{21} represents the difference in log-odds between level 2(post-treatment) and level 1(pre-treatment) and the coefficient b_{21} represents the difference in log-odds between level 2(females) and level 1(males).

Table 3.10: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	Odds ratio	95% CI for variable	95% CI for odds-ratio
Constant	a	0.338	0.111				
Time	c_{21}	-0.234	0.094	-2.49	0.79	(-0.41,-0.05)	(0.66,0.95)
Sex	b_{21}	0.201	0.097	2.07	1.22	(0.01,0.39)	(1.01,1.48)

From the fitted model in Akeredolu, we can interpret what is shown in the graph 5. There is an evidence for effect of treatment but this reduction in prevalence of ascariasis does not differ significantly among the age groups. However, there is a significant sex effect, with females to be 1.22 times more likely to be infected than males.

In Alakowe where the treatment was provided three times a year, a significant reduction in prevalence for the school-age children can easily be noticed from the graph 3.5. Table 3.11 shows the fitted models.

Table 3.11: Fitted models to data from Alakowe

Model	-2Log-likelihood	No. of parameters
Age+Sex	393.0749	4
Time+Age+Sex	364.066	5
Age+Time	369.0361	4
Time+Age+Time*Age	357.4371	6

The one which describes the data accurately is:

log-odds of the disease= $Intercept + Time + Age + Time * Age$.

The estimated coefficients and the odds-ratios along with their 95% CI's are shown in the Table 3.12.

Table 3.12: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	Odds ratio	95% CI for variable	95% CI for odds-ratio
Constant	a	-0.16	0.227				
Time	c_{21}	-0.785	0.227	-3.448	0.46	(-1.23,-0.35)	(0.29,0.70)
Age(1)	b_{21}	-0.436	0.177	-2.453	0.65	(-0.79,-0.09)	(0.45,0.91)
Age(2)	b_{31}	-0.129	0.087	-1.47	0.88	(-0.30,0.04)	(0.74,1.04)
Time*Age(1)	f_{21}	-0.37	0.177	-2.101	0.69	(-0.72,-0.02)	(0.49,0.98)
Time*Age(2)	f_{31}	0.215	0.087	2.45	1.24	(0.05,0.39)	(1.05,1.48)

In fact, there is a significant reduction in children as a result of providing treatment in four-monthly intervals. The odds of the disease is 0.69 for this age group which means that children are 0.69 times less likely to be infected by Ascaris after the effect of the treatment. However, a significant increase has been noticed in adults with an odds-ratio of 1.24 for this period of time, a result that was expected since no treatment was provided in this age group.

Finally, we will fit a model for the control village Ladin. Again several model were fitted as table 3.13 shows.

Table 3.13: Fitted models to data from Ladin

Model	-2Log-likelihood	No. of parameters
Age	440.002	3
Age+Sex	437.3027	4
Age+Time	439.9409	4
Time+Age+Sex	437.273	5

The final model is: log-odds= *Intercept* + *Age* with estimated coefficients, odds-ratios and 95% CI's as shown in table 3.14.

Table 3.14: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	Odds ratio	95% CI for variable	95% CI for odds-ratio
Constant	a	0.628	0.154				
Age(1)	b_{21}	0.169	0.185	0.912	1.18	(-0.19,0.53)	(0.83,1.70)
Age(2)	b_{31}	-0.194	0.081	-2.377	0.82	(-0.35,-0.03)	(0.70,0.97)

As we expected, there is an increased risk in prevalence for children, which does not reach statistical significance. On the other hand, adults seem to be less exposed to the disease. These results do not change for the period following the treatment since Ladin acted as a control village.

To conclude, the highest reduction in prevalence of ascariasis was noticed in Alakowe, where targeted chemotherapy was provided in children aged 5 to 15 years three times a year, while there was no difference between males and females. The treatment has also been effective in Akeredolu in which the drug levamisole was given twice a year, but no significant difference among age groups was noticed. Also, females were still more infected than males. On the other hand, the results showed that providing mass chemotherapy in China reduces the prevalence of ascariasis in all the age groups but not as much as in Nigeria. Especially in children which is the high risk group, the proportion infected drops significantly but it is still higher than that in Akeredolu and Alakowe. Moreover, females are exposed even more than males. Therefore, in order to control the prevalence of the disease in children for both males and females, we would suggest to provide targeted chemotherapy with the drug levamisole at four-monthly intervals.

Intensity

The next step is to carry out an analysis for the intensity of ascariasis. Unlike prevalence, in this case we focus only on the number of eggs per infected person assuming that observations with zero eggs are missing. This allows us to model the intensity of eggs conditional on the presence of eggs. However, this generates even more missing data and in the case of a continuous scale Azzalini's model can not be fitted. For this reason, we will fit a linear model from the gaussian family with independent Normal errors.

Model for intensity

Assume that Y is a random variable whose components are independent Normal variables with constant variance σ^2 and $E(Y) = \mu$ where $\mu = X\beta$, where X is the nxp model matrix, n is the number of observations and p the number of parameters. Then the generalized linear model for intensity will have the form:

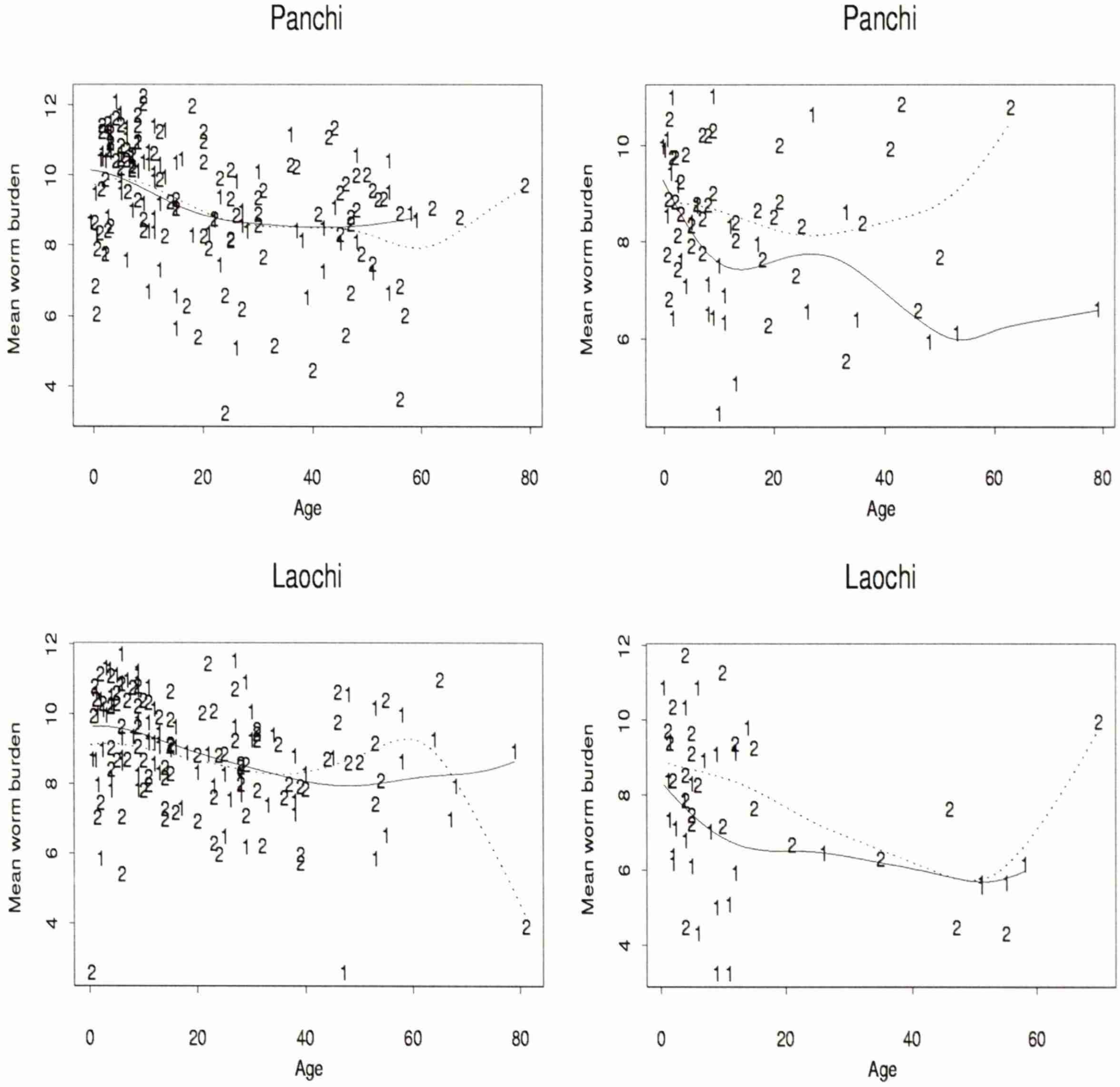
$$\text{log}E(Y) = \mu = x^T\beta$$

where β is a vector of the parameters of interest and x^T is the matrix of covariates $x_1, x_2, ..., x_p$. So, β describes the change in the log of the population average count per unit change in the explanatory variable. For example, if β is the coefficient associated with age and time then a negative value of β is evidence that the treatment is effective for this age group and period of time. A log transformation of the response is used to stabilize the variance over time.

Fitting the model

Using the linear model described above, we will estimate the covariate effects of time, age and sex in the intensity of ascariasis. In the following graph, we can roughly see the intensity of ascariasis in Panchi and Laochi just before and immediately after the treatment. The number of eggs has been transformed in a more amenable scale, i.e. the logarithmic scale. To produce the curves we use the non-parametric regression method described in section 2.4a. Again, we have to notice that because for our data the aim is descriptive, we use a value of bandwidth $h = 10$ which has been chosen subjectively. As it is apparent from the graph 3.6, the intensity seems to have a similar trend to prevalence. It is higher in children but lower in adults with slightly higher values for females.

Figure 3.6: Smooth curves of Intensity of ascariasis to age in China. Left:Pre-treatment, Right:Post-treatment



Following exactly the same procedure as we did for prevalence(i.e. age is considered as a factor with 3 levels, sex as a factor with 2 levels and time as a factor with 11 levels), several models have been fitted to the data, starting with Panchi. These are shown in table 3.15.

Table 3.15: Fitted models

Model	**-2Log-likelihood**	**No. of parameters**
Age+Sex	4342.214	4
Time+Age+Sex	4075.148	14
Time+Age+Sex+Age*Sex	4028.224	16
Time+Age+Sex+Age*Sex+Time*Age	3945.64	36

We end up with the model:

Intensity= $a + Time + Age + Sex + Time * Age + Age * Sex$, where Intensity corresponds to the log(eggs+1), Time*Age and Age*Sex are the interaction effects between Time, Age and Age, Sex respectively.

Table 3.16 shows the estimated coefficients along with the corresponding 95% CI's. Similar to the model fitting procedure for prevalence, we assume that age is a factor with 3 levels in order to describe in a more accurate way in which group the intensity of the disease drops as well as the size of this reduction. Also, time is again a factor with 11 levels (6 before and 5 after treatment) but here we present only the significant one, which corresponds to the period just immediately after treatment. The coefficients f_{21} and f_{31} associate age and time(i.e. children aged 0 to 15, 5 to 15 years and adults for the post-treatment period), g_{21} and g_{31} associate age and sex(i.e. children aged 0 to 5, 5 to 15 years and adults for males and females).

Table 3.16: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	95% CI for variable
Constant	a	8.804	0.046		
Time	c_{76}	-0.089	0.029	-3.077	(-0.15,-0.03)
Age(1)	b_{21}	-0.187	0.06	-3.127	(-0.31,-0.07)
Age(2)	b_{31}	-0.528	0.03	-17.23	(-0.59,-0.47)
Sex	d_{21}	0.159	0.044	3.58	(0.07,0.25)
Time*Age(1)	f_{21}	-0.156	0.074	-2.11	(-0.31,-0.01)
Time*Age(2)	f_{31}	-0.034	0.015	-2.25	(-0.06,-0.004)
Sex*Age(1)	g_{21}	0.209	0.058	3.565	(0.10,0.32)
Sex*Age(2)	g_{31}	0.06	0.028	2.12	(0.01,0.11)

As we can see from table 3.16, the effect of the treatment reduces significantly the intensity not only in adults but also in children(there are negative values for both coefficients f_{21} and f_{31}), while females still remain the high risk group(there are positive values for g_{21} and g_{31}).

As for the intensity in Laochi, the fitted models are shown in table 3.17.

Table 3.17: Fitted models

Model	-2Log-likelihood	No. of parameters
Age+Sex	4350.84	4
Time+Age+Sex	4091.971	14
Time+Age+Sex+Age*Time	4004.526	34
Time+Age+Sex+Age*Sex+Time*Age	3967.602	36
Time+Age+Sex+Age*Sex+Time*Age+Time*Sex	3916.258	46
Time*Age*Sex	3850.029	66

The model which describes the data accurately is the maximal one:

Intensity= $a+Time+Age+Sex+Time*Age+Time*Sex+Age*Sex+Time*Age*Sex$. The estimated coefficients along with the corresponding 95% CI's are shown in table 3.18. There are also coefficients which associate sex and time(g_{21}), as well as, sex, age and time(j_{21}).

Table 3.18: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	95% CI for variable
Constant	a	6.97	0.150		
Time	c_{76}	-0.239	0.036	-6.629	(-0.31,-0.17)
Age(1)	b_{21}	-0.163	0.063	-2.57	(-0.28,-0.04)
Age(2)	b_{31}	-0.616	0.033	-18.2	(-0.68,-0.56)
Sex	d_{21}	0.100	0.049	2.019	(0.00,0.20)
Time*Age(1)	f_{21}	-0.047	0.024	-1.94	(-0.10,0.10)
Time*Age(2)	f_{31}	-0.034	0.016	-2.18	(-0.06,-0.004)
Time*Sex	g_{21}	0.081	0.036	2.24	(0.01,0.15)
Age(2)*Sex	h_{21}	0.102	0.033	3.04	(0.04,0.16)
Time*Age(2)*Sex	j_{21}	0.035	0.011	2.96	(0.02,0.06)

Similar to Panchi, the intensity in Laochi reduced significantly in children and adults(negative values for f_{21} and f_{31}). However in adults, females are exposed more to the disease(positive values for h_{21} and j_{21}).

We will try now to assess the effect of the treatment in the intensity of ascariasis in Nigerian villages as well as the age and sex effects. Again, we consider age as a factor with 3 levels, sex as a factor with 2 levels and time as a factor with 2 levels. In Iyanfoworogi different models have been fitted to the data, as table 3.19 shows.

Table 3.19: Fitted models to data from Iyanfoworogi

Model	-2Log-likelihood	No. of parameters
Time+Age	696.1312	4
Time+Age+Sex	695.9483	5
Age+Time+Time*Age	693.2785	6

The one which describes the data accurately is:

Intensity= $a + Time + Age$, and the estimated coefficients with the 95% CI's are shown in table 3.20.

Table 3.20 shows that, although there is a significant effect of the treatment, similar to prevalence, the intensity in young children is not reduced significantly(negative and

Table 3.20: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	95% CI for variable
Constant	a	8.901	0.049		
Time	c_{21}	-0.625	0.123	-5.051	(-0.87,-0.39)
Age(1)	b_{21}	0.020	0.159	0.126	(-0.29,0.33)
Age(2)	b_{31}	-0.177	0.085	-2.059	(-0.35,-0.01)

positive values for c_{21} and b_{21} respectively).

In Akeredolu, the fitted models are as shown in table 3.21.

Table 3.21: Fitted models to data from Akeredolu

Model	-2Log-likelihood	No. of parameters
Time+Age	944.3398	4
Time+Age+Sex	942.7698	5
Age+Time+Time*Age	943.4602	6

We end up to a similar model with the data from the previous village with estimated coefficients as are shown in table 3.22.

Table 3.22: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	95% CI for variable
Constant	a	6.99	0.118		
Time	c_{21}	-0.402	0.113	-3.55	(-0.62,-0.18)
Age(1)	b_{21}	0.343	0.152	2.249	(0.039,0.647)
Age(2)	b_{31}	-0.294	0.078	-3.777	(-0.44,-0.14)

Similar to prevalence, the intensity in Akeredolu did reduce significantly due to the effect of treatment, but this reduction did not reach significant levels in children(positive value for b_{21}).

In Alakowe, where targeted chemotherapy was provided in children three times a year, the fitted models to these data are shown in table 3.23.

Table 3.23: Fitted models to data from Alakowe

Model	-2Log-likelihood	No. of parameters
Time+Age	496.6664	4
Age+Sex	494.7511	4
Time+Age+Sex	489.9484	5
Age+Time+Sex+Time*Age	473.8214	7
Age+Time+Sex+Time*Age+Age*Sex	470.1198	9
Age+Time+Sex+Time*Age+Time*Sex	472.4557	8

The model which describes the data accurately is: Intensity= $a + Time + Age + Sex + Time * Age$, which means that there are significant differences in intensity among the age groups for the period after the treatment although there are not individually significant as we can see in table 3.24.

Table 3.24: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	95% CI for variable
Constant	a	6.327	0.373		
Time	c_{21}	-0.660	0.349	-1.88	(-1.35,0.03)
Age(1)	b_{21}	-0.116	0.272	-0.42	(-0.65,0.41)
Age(2)	b_{31}	-0.112	0.133	-0.83	(-0.37,0.15)
Sex	d_{21}	0.184	0.188	0.979	(-0.19,0.55)
Time*Age(1)	f_{21}	-0.205	0.259	-0.789	(-0.72,0.30)
Time*Age(2)	f_{31}	0.186	0.132	1.407	(-0.07,0.45)

It is apparent from the above table, the reduction in intensity of ascariasis in children(negative value for f_{21}). However, this reduction did not reach significant levels (95% CI:(-0.72, 0.30)).

Finally, in the control village Ladin, the following models have been fitted.

Table 3.25: Fitted models to data from Ladin

Model	-2Log-likelihood	No. of parameters
Time+Age	736.2726	4
Time+Sex	732.8659	3
Age+Sex	761.1758	4
Time+Age+Sex	717.9495	5
Age+Time+Sex+Time*Age	701.4452	7

The one which describes the data properly is: Intensity= $a + Time + Age + Sex + Time * Age$. The estimated coefficients with the 95% CI's are shown in table 3.26.

Table 3.26: Fitted model parameters

Variable	Term	Coeff.	Std. error	t value	95% CI for variable
Constant	a	6.656	0.365		
Time	c_{21}	-0.325	0.191	-1.69	(-0.70,0.05)
Age(1)	b_{21}	-0.401	0.237	-1.69	(-0.86,0.06)
Age(2)	b_{31}	-0.246	0.106	-2.316	(-0.46,-0.04)
Sex	d_{21}	0.424	0.651	0.652	(-0.85,1.69)
Time*Age(1)	f_{21}	-0.509	0.237	-2.14	(-0.97,-0.05)
Time*Age(2)	f_{31}	-0.142	0.106	-1.337	(-0.35,0.07)

Although no treatment was provided in this village, the intensity was reduced significantly in children but not in adults. However, as we saw above, the proportion of children who have been infected by the disease did not drop significantly.

To sum up, we can say that mass chemotherapy treatment provided in China caused a significant drop in intensity of ascariasis in children and adults. On the other hand, providing targeted chemotherapy twice and three times a year, we achieve to reduce the intensity but not significantly. It is unclear why a reduction in Ascaris intensity was observed in the control village, which received no treatment. This may be explained by assuming either a genuine disruption in transmission of Ascaris or a stimulation in the community to seek help and medication on seeing the beneficial effects in the children.

3.4 Conclusions

In summary, we can say that providing mass chemotherapy treatment in China had the result of reducing the prevalence of ascariasis not only in adults but also in children who are more exposed to the disease. This reduction in prevalence is followed by a corresponding reduction in intensity. However, females not only harbour more worms than males but also are infected more by them. This trend is more obvious in adults and it is expected since women in this rural province of China are more exposed to the contaminated soil than men.

On the other hand, providing targeted chemotherapy in children with different frequencies, has the result of reducing not only the prevalence but also the intensity of ascariasis. Specifically, the highest reduction in prevalence was noticed in Alakowe where the treatment was provided at four-monthly intervals. However, the treatment had no effect when provided once a year. Although, the intensity in Akeredolu and Alakowe was reduced after treatment, there was no significant difference among those who received treatment and those who did not.

From these results we conclude that mass chemotherapy treatment reduces significantly not only the prevalence but also the intensity of ascariasis in children and adults, but there is still a significant difference between males and females. As for the targeted chemotherapy with the drug levamisole, this reduces significantly the prevalence of ascariasis especially when the drug is provided on four-monthly intervals. It seems to be more effective than the mass chemotherapy because children who have been treated with this type of treatment are less likely to be infected by the disease than those treated with the latter. However, it does not seem to have analogous effects for reducing the intensity of the disease. In the following chapter, we will estimate the size of the treatment effect in prevalence and intensity between mass and targeted chemotherapy as well as the time needed to recover to pre-treatment levels as a result of re-infection.

SECTION FOUR

Analysis of treatment effects

Chapter 4

Analysis of treatment effects

4.1 Introduction

So far, we have seen that providing either mass or targeted chemotherapy treatment, we are able to control prevalence and also the intensity of ascariasis in regions where the disease is widely spread. Specifically, there is significant evidence that the former reduces not only the prevalence but also the intensity in both children and adults. As for the latter, it seems to be more effective when it is applied three times a year, at four-monthly intervals. Therefore, the questions arising are how big this treatment effect is for each village separately and the presence and size of any differences and how big is such a difference among the villages. In other words, our primary interest in this chapter is to estimate the size of the effect of the treatment for each age group, make comparisons among the villages and then investigate the required time for prevalence and intensity to recover to the previous value as a result of re-infection.

4.2 Comparisons within each village

Prevalence

In order to estimate the size of the effect of the treatment in each village, we take two samples, one just before and another one just immediately after the treatment and estimate a 95% confidence interval for the true paired difference in proportions. This 'before' and 'after' design includes the proportions p_{11}, p_{10}, p_{01} and p_{00} for those who are infected before and after treatment, those who are infected before but not after treatment, those who are not infected before but are infected after treatment and those who are not infected neither before nor after treatment respectively. We concentrate on the p_{10} and p_{01} since the proportions p_{11} and p_{00} give no indication for treatment difference. Also, the difference in the marginal proportions of before and after is $(p_{11}+p_{10})-(p_{01}+p_{11}) = p_{10}-p_{01}$. Hence, the variance of the estimated difference $\hat{p}_{10}-\hat{p}_{01}$ will be:

$$var(\hat{p}_{10}-\hat{p}_{01}) = var(\hat{p}_{10}) + var(\hat{p}_{01}) - 2cov(\hat{p}_{10}, \hat{p}_{01})$$

where $cov(\hat{p}_{10}, \hat{p}_{01})$ denotes the covariance between $\hat{p}_{10}$ and $\hat{p}_{01}$.

After some algebra we obtain:

$$var(\hat{p}_{10}-\hat{p}_{01}) = \frac{p_{10}+p_{01}-(p_{10}-p_{01})^2}{n}$$

where n is the total number of observations. Then the estimated standard error will have the form:

$$s.e.(\hat{p}_{10}-\hat{p}_{01}) = \sqrt{\frac{p_{10}+p_{01}-(p_{10}-p_{01})^2}{n}}$$

and hence a $100(1-\alpha)\%$ interval estimate for the difference will be:

$$\hat{p}_{10} - \hat{p}_{01} \pm z_{a/2} s.e.(\hat{p}_{10} - \hat{p}_{01})$$

(G. W. Snedecor, W. G. Cohran(1980)). So, for Panchi a 95% confidence interval for the difference in the proportions infected by ascariasis was (0.32, 0.51) which does not contain 0.0 and hence there is a significant evidence for reduction in prevalence of the disease. Table 4.1 shows the proportions infected by ascariasis in each age group of Panchi village along with the corresponding 95% confidence intervals.

Table 4.1: Confidence intervals for the difference between pre- and post-treatment values by age in Panchi

Age	**Pre-treatm.**	**Post-treatm.**	**95% CI**
0 - 5	0.40	0.12	(0.06, 0.50)
5 - 15	0.50	0.05	(0.27, 0.63)
>15	0.52	0.05	(0.35, 0.59)

As it can be shown from the above table, the highest reduction was observed in children aged 5 to 15 years and adults. Analogously, we present the 95% CI's for the difference in proportions before and just immediately after treatment in males and females, as it is indicated in table 4.2.

Table 4.2: Confidence intervals for the difference between pre- and post-treatment values by sex in Panchi

Sex	**Pre-treatm.**	**Post-treatm.**	**95% CI**
Males	0.47	0.08	(0.26, 0.52)
Females	0.5	0.046	(0.33, 0.57)

It is apparent the significant reduction in prevalence for both males and females without any remarkable difference between them.

Similarly to Panchi, the prevalence in Laochi was reduced significantly just immediately after the treatment and the difference between pre and post-treatment lies in the interval (0.34, 0.54). The interval estimates for the difference in proportions for each age group as well as the corresponding ones for males and females are shown in Tables 4.3 and 4.4 respectively.

Table 4.3: Confidence intervals for the difference between pre- and post-treatment by age in Laochi

Age	**Pre-treatm.**	**Post-treatm.**	**95% CI**
0 - 5	0.30	0.03	(0.09, 0.45)
5 - 15	0.53	0.00	(0.37, 0.69)
>15	0.51	0.04	(0.35, 0.59)

Table 4.4: Confidence intervals for the difference between pre- and post-treatment values by sex in Laochi

Sex	**Pre-treatm.**	**Post-treatm.**	**95% CI**
Males	0.44	0.026	(0.29, 0.53)
Females	0.5	0.028	(0.34, 0.59)

Again, the highest reduction was observed in young children and adults. Specifically, the prevalence was reduced dramatically from 37% to 69% in children aged 5 to 15 years and from 35% to 59% in adults. Similarly to Panchi, the prevalence dropped significantly in both sexes.

For the Nigerian villages in which targeted chemotherapy treatment was provided only in children aged 5 to 15 years and the time interval between pre- and post-treatment was one year, a significant reduction in prevalence was noticed only in Alakowe. Table 4.5 shows the difference between pre- and post-treatment values only for school-age children before and after treatment while Table 4.6 shows the corresponding differences in males and females along with the 95% confidence intervals for each village.

Table 4.5: Confidence intervals for the difference between pre- and post-treatment by village for school-age children in Nigeria

Village	**Pre-treatm.**	**Post-treatm.**	**95% CI**
Iyanfoworogi	0.11	0.21	(-0.32, 0.12)
Akeredolu	0.22	0.08	(-0.04, 0.32)
Alakowe	0.58	0.03	(0.35, 0.75)
Ladin	0.18	0.23	(-0.32, 0.22)

Table 4.6: Confidence intervals for the difference between pre- and post-treatment by village and sex in Nigeria

Village	**Sex**	**Pre-treat**	**Post-treat**	**95% CI**
Iyanfoworogi	Males	0.058	0.35	(-0.56, -0.02)
	Females	0.17	0.11	(-0.19, 0.31)
Akeredolu	Males	0.19	0.077	(-0.08, 0.3)
	Females	0.3	0.15	(-0.2, 0.5)
Alakowe	Males	0.6	0.0	(0.39, 0.81)
	Females	0.53	0.058	(0.2, 0.74)
Ladin	Males	0.14	0.14	(-0.39, 0.39)
	Females	0.0	0.5	(-0.89, -0.11)

In fact, providing targeted chemotherapy in children three times a year with the drug levamisole(i.e. in Alakowe), the prevalence was reduced significantly from 35 to 75%. Also in Akeredolu where the drug was provided twice a year, a reduction was noticed but it did not reach significant levels. On the other hand, the proportion infected by Ascaris in Iyanfoworogi and the control village Ladin increased, although not significantly. Also, significant reduction was observed only in Alakowe for both sexes and especially in males where the prevalence dropped dramatically from 60% to approximately 0%. In none of the other groups was an analogous reduction noticed. In contrast, the prevalence was increased significantly in males from Iyanfoworogi and females from Ladin.

Intensity

Similarly to prevalence, the intensity in Panchi was reduced significantly just immediately after treatment(i.e. paired t-tests reveal these differences) in all age groups for males and females, as tables 4.7 and 4.8 show respectively.

Table 4.7: Difference in pre- and post-treatment intensity by age in Panchi

Age	Pre-treatm.	Post-treatm.	95% CI
0-5	3.25	1.77	(0.57, 2.36)
5-15	3.82	1.55	(1.6, 2.93)
> 15	2.31	0.51	(1.33, 2.26)

Table 4.8: Difference in pre- and post-treatment intensity by sex in Panchi

Sex	Pre-treatm.	Post-treatm.	95% CI
Males	2.69	0.96	(1.19, 2.25)
Females	3.11	1.17	(1.46, 2.42)

Analogously, the number of worms per person was reduced significantly in Laochi not only for all the age groups but also for both sexes, as tables 4.9 and 4.10 show.

Table 4.9: Difference in pre- and post-treatment intensity by age in Laochi

Age	Pre-treatm.	Post-treatm.	95% CI
0-5	3.88	2.49	(0.63, 2.15)
5-15	3.49	0.91	(1.93, 3.22)
> 15	2.12	0.24	(1.44, 2.32)

It is apparent from these tables that the highest reduction in intensity of ascariasis was noticed in children aged 5 to 15 years, while it does not seem to be a big difference between males and females. In the Nigerian villages, the impact of treatment is significant only in Alakowe as is shown in Table 4.11.

In this village, the size of the effect of the treatment lies in a 95% confidence interval of (1.35, 2.51), which does not contain 0.0, so there is a significant evidence that there is a

Table 4.10: Difference in pre- and post-treatment intensity by sex in Laochi

Age	Pre-treatm.	Post-treatm.	95% CI
Males	2.73	0.86	(1.42, 2.3)
Females	2.9	0.85	(1.55, 2.54)

Table 4.11: Difference in pre- and post-treatment intensity for school-age children in Nigeria

Village	Pre-treatm.	Post-treatm.	95% CI
Iyanfoworogi	2.01	1.81	(-0.54, 0.93)
Akeredolu	1.96	1.42	(-0.05, 1.15)
Alakowe	2.4	0.46	(1.35, 2.51)
Ladin	2.4	2.08	(-0.47, 1.10)

reduction in intensity for children aged 5 to 15 years especially when the drug is provided at four-monthly intervals. If we now estimate the treatment effect in each village for males and females of the same age group (i.e. aged 5 to 15 years), then a corresponding reduction was observed only in the same village (i.e. Alakowe), although a remarkable reduction was also noticed in females from Iyanfoworogi, as it is indicated from table 4.12.

Table 4.12: Difference in pre- and post-treatment intensity by sex in Nigeria

Village	Sex	Pre-treat	Post-treat	95% CI
Iyanfoworogi	Males	1.24	1.85	(-1.8, 0.57)
	Females	2.77	1.76	(0.19, 1.8)
Akeredolu	Males	1.65	1.36	(-0.45, 1.04)
	Females	2.28	1.27	(-0.29, 2.31)
Alakowe	Males	2.67	0.69	(1.2, 2.76)
	Females	2.05	0.19	(0.88, 2.82)
Ladin	Males	3.05	2.47	(-1.44, 2.61)
	Females	1.8	2.37	(-2.15, 1.03)

4.3 Comparisons between the countries

We now move on to another comparison, this time among the villages of both countries. In order to compare the size of the effect of the treatment between Chinese and Nigeria villages, we have to take samples within the same period of time. Thus, we take the proportions infected by ascariasis in both countries one year after the treatment provided and test whether these proportions differ or not. Also, we focus only on the children aged 5 to 15 years since this is the group with the highest risk of infection in both males and females.

When there are two independent samples with sizes, say n_1 and n_2 with number of successes y_1 and y_2 respectively then the corresponding probabilities of success will be $\hat{p}_1 = \frac{y_1}{n_1}$ and $\hat{p}_2 = \frac{y_2}{n_2}$. So, the variance of the estimated difference between $\hat{p}_1$ and $\hat{p}_2$ will have the form:

$$var(\hat{p}_1 - \hat{p}_2) = var(\hat{p}_1) + var(\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

and the standard error:

$$s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Hence a $100(1-\alpha)\%$ CI for the difference between the two proportions is:

$$\hat{p}_1 - \hat{p}_2 \pm z_{a/2} s.e.(\hat{p}_1 - \hat{p}_2)$$

In our case, we obtain interval estimates for the difference in one-year post-treatment values between Chinese and Nigeria villages. The results are shown in Table 4.13.

As it can be shown from this table, there is no evidence for significant difference between Chinese and Nigeria villages for the proportion infected by the disease after a one year

Table 4.13: Interval estimates for the difference in one year post-treatment values among the villages for males and females

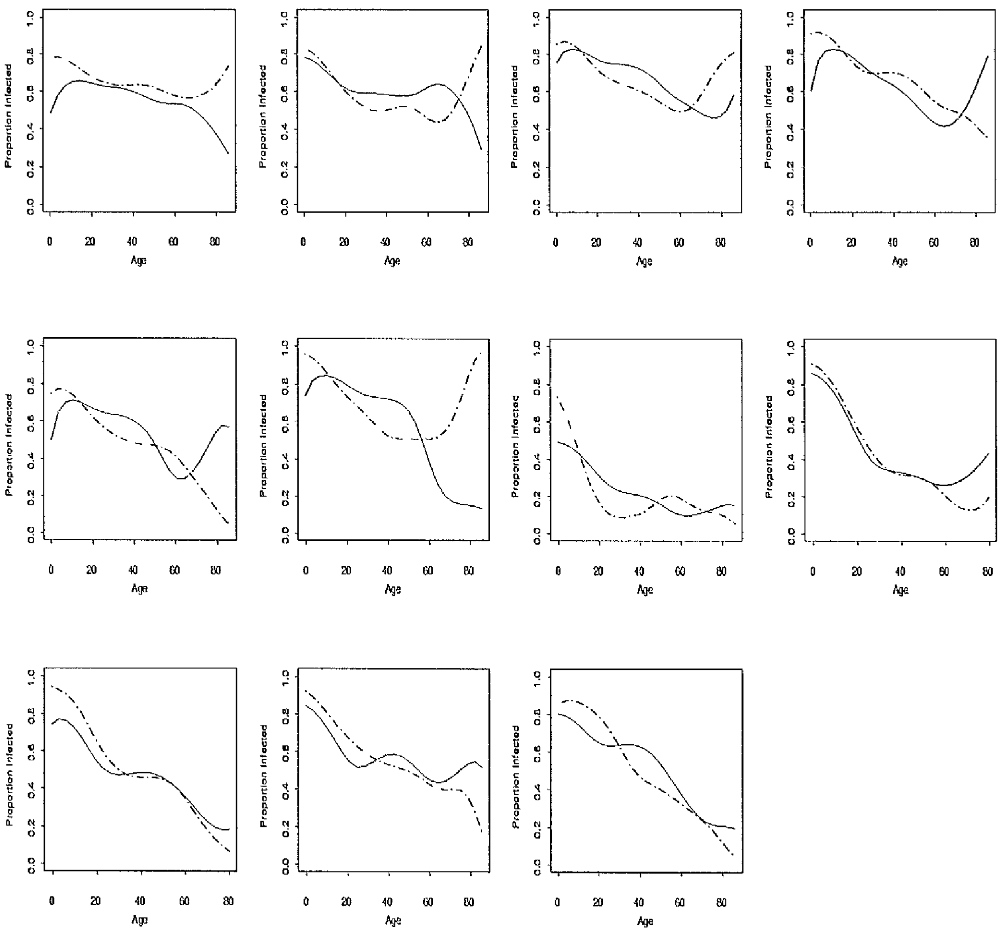
Mass Chemotherapy	Targeted Chemotherapy	Sex	95% CI
Panchi	Iyanfoworogi	Males	(-0.528, -0.068)
		Females	(-0.235, 0.157)
Panchi	Akeredolu	Males	(-0.165, 0.115)
		Females	(-0.3, 0.15)
Panchi	Alakowe	Males	(-0.047, 0.15)
		Females	(-0.157, 0.18)
Laochi	Iyanfoworogi	Males	(-0.53, -0.073)
		Females	(-0.136, 0.256)
Laochi	Akeredolu	Males	(-0.16, 0.1)
		Females	(-0.21, 0.25)
Laochi	Alakowe	Males	(-0.041, 0.135)
		Females	(-0.068, 0.292)

period treatment. Although, there are some exceptions such as the difference between Chinese villages and Iyanfoworogi where mass treatment seems to be more effective in males, we could generally admit that the two types of treatment have similar effects in controlling the prevalence of the disease.

4.4 Re-infection

Our aim in this section is to investigate the time required for the prevalence to recover to the previous value, that is we want to estimate how long the effect of the treatment lasts for and when the proportion infected by the disease reaches the peak value as a result of re-infection. To achieve this, we focus only on the Chinese villages since we have taken samples in two monthly intervals only from these villages. The following graph shows the proportion infected by ascariasis in each village of China over time.

Figure 4.1: Prevalence of ascariasis in Panchi(Continuous line) and Laochi(Dashed line)



The first six graphs correspond to the period before and the other five to the period following the mass chemotherapy treatment. As we can clearly see, in the period just before the treatment, the prevalence was extremely high in both villages especially in young children but lower in adults. Specifically, it was more than 80% for children in Panchi and more than 90% for the same age group in Laochi. Just immediately after the treatment, it drops dramatically not only in children but also in adults. However, as is shown roughly from the graphs, young children are reinfected earlier than adults. In order to have a formal conclusion about the period of the re-infection, we will estimate 95%

confidence intervals for the difference in proportions for the before and after treatment measures, as it was described in section 4.2. In this case, the prevalence in Panchi, reached the peak (80%) at six months after treatment in children aged 0 to 5 years, at nine months in school-age children and at eleven months in adults, as it is shown in the following table.

Table 4.14: Recovery times for prevalence and 95% CI's for the difference between pre- and post-treatment by age in Panchi

Age	Period of Re-infection	95% CI
0 - 5	June - December	(-0.22, 0.12)
5 - 15	June - March	(-0.1, 0.22)
> 15	June - May	(-0.02, 0.26)

Similarly to Panchi the prevalence in Laochi reached the previous levels in the same time intervals except the second age group in which the proportion infected earlier than in Panchi. The 95% confidence intervals for the difference in proportions for before and after treatment in Laochi are shown in the following table.

Table 4.15: Recovery times for prevalence and 95% CI's for the difference between pre- and post-treatment by age in Laochi

Age	Period of Re-infection	95% CI
0 - 5	June - December	(-0.014, 0.286)
5 - 15	June - December	(-0.05, 0.23)
> 15	June - May	(-0.08, 0.22)

As we can see from the above tables, all the interval estimates contain 0.0, so there is no evidence for significant difference between pre- and post-treatment values for the proportion infected by Ascaris.

Similarly to prevalence, the intensity in Panchi reached the previous level at six months post-treatment in children aged 0 to 5 years and at nine months in school-age children, except from adults, where it reaches the previous value at the end of the study in July '95, as the Table 4.16 shows.

As for the second village Laochi, the re-infection rates in intensity are shown in Table

Table 4.16: Re-Infection in Panchi(intensity)

Age	**Pre-treatm.**	**Post-treatm.**	**95% CI**
0-5	3.28	3.32	(-0.82, 0.74)
5-15	3.79	3.42	(-0.36, 1.11)
> 15	2.18	1.98	(-0.27, 0.67)

4.17.

Table 4.17: Re-Infection in Laochi(intensity)

Age	**Pre-treatm.**	**Post-treatm.**	**95% CI**
0-5	3.84	3.52	(-0.44, 1.087)
5-15	3.49	2.94	(-0.08, 1.19)
> 15	2.18	1.73	(-0.11, 1.0026)

It is apparent from the above table that similar to prevalence, the intensity in Laochi reaches the previous level at six and eleven months after treatment in children and adults respectively. Again, there is no remarkable difference between pre- and post-treatment values in intensity, as it is indicated from the interval estimates.

4.5 Conclusions

To sum up, providing mass chemotherapy treatment in Panchi and Laochi reduces significantly the prevalence and intensity of ascariasis for all the age groups and especially in young children and adults for both sexes. The period of re-infection fluctuates from six to eleven months in children and adults respectively. On the other hand, targeted chemotherapy in children reduces significantly both the prevalence and intensity of the disease but only if it is provided at four-monthly intervals. However, no significant difference in prevalence was noticed between mass and targeted chemotherapy. The last result is in contrast with the one from the previous chapter in which the fitted linear model suggests a non-significant reduction in intensity for the treated group in Alakowe. Thus, in the following chapter, we will investigate the effect of age as a linear predictor for this data set firstly when the response variable is binary and secondly when the response is a continuous variable.

SECTION FIVE

Generalized Additive Models

Chapter 5

Generalized Additive Models

5.1 Introduction

In this Chapter, we make use of generalized additive models in order to incorporate non-linearities in a more direct way into a model. Such models are useful in terms that they provide a means of checking more formally on linearity assumptions and also can provide a way of modelling the data even when these are non-linear but effects can be assumed to be smooth. However, generalized additive models have two main disadvantages. The first one is that they do not include correlation across time. The second disadvantage is that even when we assume independence, the distributional results about chi-square distributions in the model comparisons are actually not correct. Hence any inference drawn from such a model is approximate.

A generalized additive model is a generalization of a linear model. The only difference is that an additive predictor $\eta = \alpha + \sum_j f_j(x_j)$ replaces the linear predictor $\eta = \sum_j \beta_j$. The functions $f_j(x_j)$ are smooth functions estimated from the data.

There are two algorithms for fitting such a model. In the first one, we consider the fitting procedure for a single covariate x. That is, we associate a set of neighbours on the x axis with each point (y_i, x_i) and estimate the value of the response function at x_i for each i by applying the standard generalized linear model algorithm with linear predictor $\alpha+\beta x_i$ in that neighbourhood. So $f_j(x_i)$ is the fitted value of $\eta-\alpha$ at x_i. This algorithm is called the local-scoring algorithm and is equivalent to using a weighted running-lines smoother on the dependent variable. Its usage is efficient because we can easily update the running-lines smoother as we pass from one neighbourhood to the next.

The second algorithm is called the back-fitting algorithm and it has the feature that it cycles through the individual terms in the additive model and updates each using an appropriate smoother. It does this by smoothing suitably defined partial residuals. Specifically, we distinguish two cases:

a) The response variable is continuous

Assume that we have a response variable y and two covariates x_1 and x_2. Then the additive model which describes the relationship between dependent and independent variables will have the form:

$$y_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \epsilon_i, i = 1, 2, ..., n$$

where f_1 and f_2 are smooth functions which sum up to zero so that $\sum_{j=1}^{n} f_j(x_{ij}) = 0$ for each i.

Then in order to construct estimates for each component, we proceed as follows:

i) Firstly, we estimate the intercept parameter β_0 by taking the mean of the responses $y^{'}$ with respect to the restrictions on each additive component to sum to zero over the observed covariate values.

ii) As a starting point, we take an estimate of the component f_2. This is provided by a non-parametric regression of the response variable y on the covariate x_2. It can be written as $\hat{f}_2 = S_2(y - y^{'})$ where S_2 is the smoothing matrix.

iii) A rearrangement of our model then is:

$$y_i - \beta_0 - f_2(x_{2i}) = f_1(x_{1i}) + \epsilon_i$$

which suggests that an estimate of the component f_1 can then be obtained by smoothing the residuals of the data after fitting $\hat{f}_2$, as follows: $\hat{f}_1 = S_1(y - y' - \hat{f}_2)$

Similarly, we can take new estimates of f_2 as: $\hat{f}_2 = S_2(y - y' - \hat{f}_1)$. In the case where more than two covariates are involved, an iterative process can be applied to produce joint estimates of each component. This is called backfitting by Buja et al.(1989) and is an extension of the Gauss-Seidel method of solving linear least squares regression problems.

It gives an estimation of each component by smoothing the residuals. Specifically, a model of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ji}) + \epsilon_i$$

can be fitted by repeated construction of the smooth estimates

$$\hat{f}_k = S_k(y - y' - \sum_{j \neq k} \hat{f}_j)$$

A complete description of the convergence properties of this algorithm is not yet available. Some results do exist, particularly where the smoothing matrices correspond to the use of smoothing splines.

Another issue arises when we want to compare such models and draw some inferences from them. Although this is not immediately obvious, Hastie and Tibshirani(1990) recommend the use of residual sum of squares and their associated approximate degrees of freedom to provide guidance for model comparisons. These are defined as:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the fitted value produced by evaluating the additive model at the observation x_i. Then each component of the fitted additive model can be represented in the form $R_j y$, for some matrix R_j, since for each step of the backfitting algorithm we get a smoothing matrix S_j. In this way, approximate degrees of freedom are calculated for each component and for the entire fitted model. In addition, because of the linear structure of the fitted model, we can compute standard errors and construct variability bands.

b) The response variable is binary

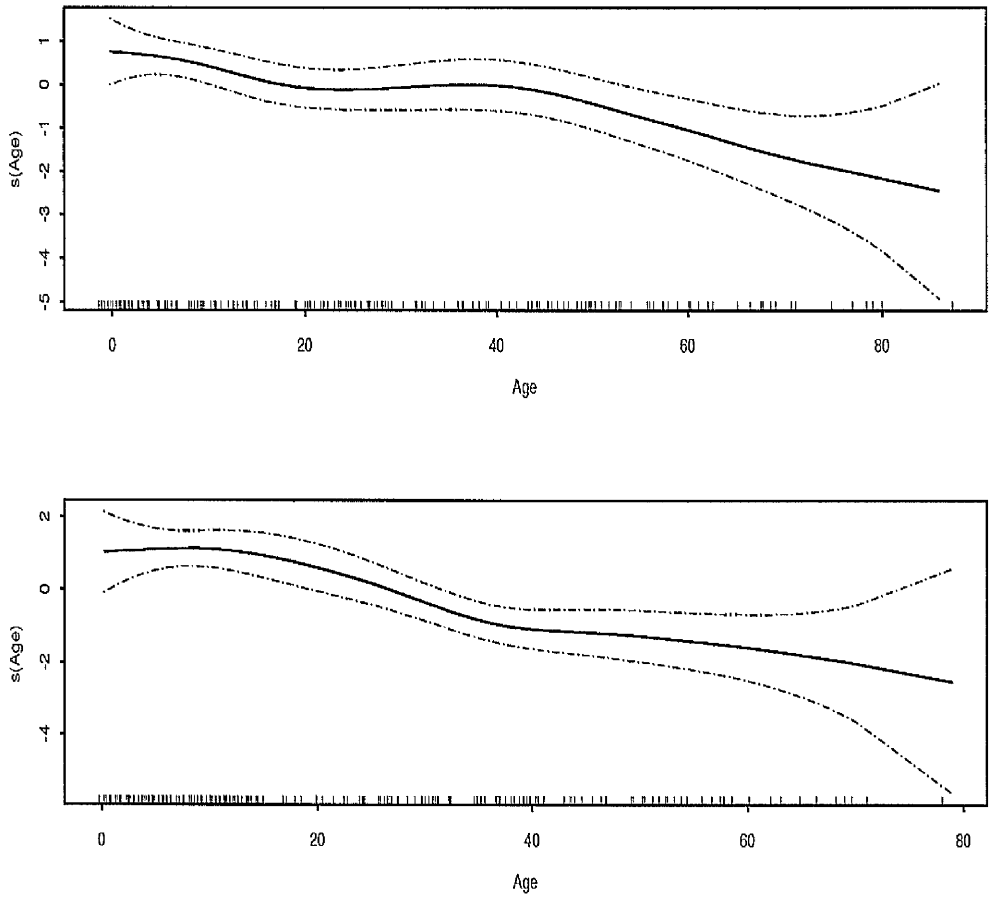
Again, for each $j = 1, 2, ..., p$ we form the residuals $r_i = z_i - \hat{\eta} + \hat{f}_j(x_{ji})$ where z_i is the dependent variable and $\hat{\eta}$ is the additive predictor. We then smooth r_i by applying a locally weighted version of the generalized linear model algorithm. In this way, we take new estimates for $\hat{f}_j(x_{ji})$ and repeat this procedure until the deviance stabilizes. In this chapter, we will make use of the backfitting algorithm.

5.2 Results

Prevalence

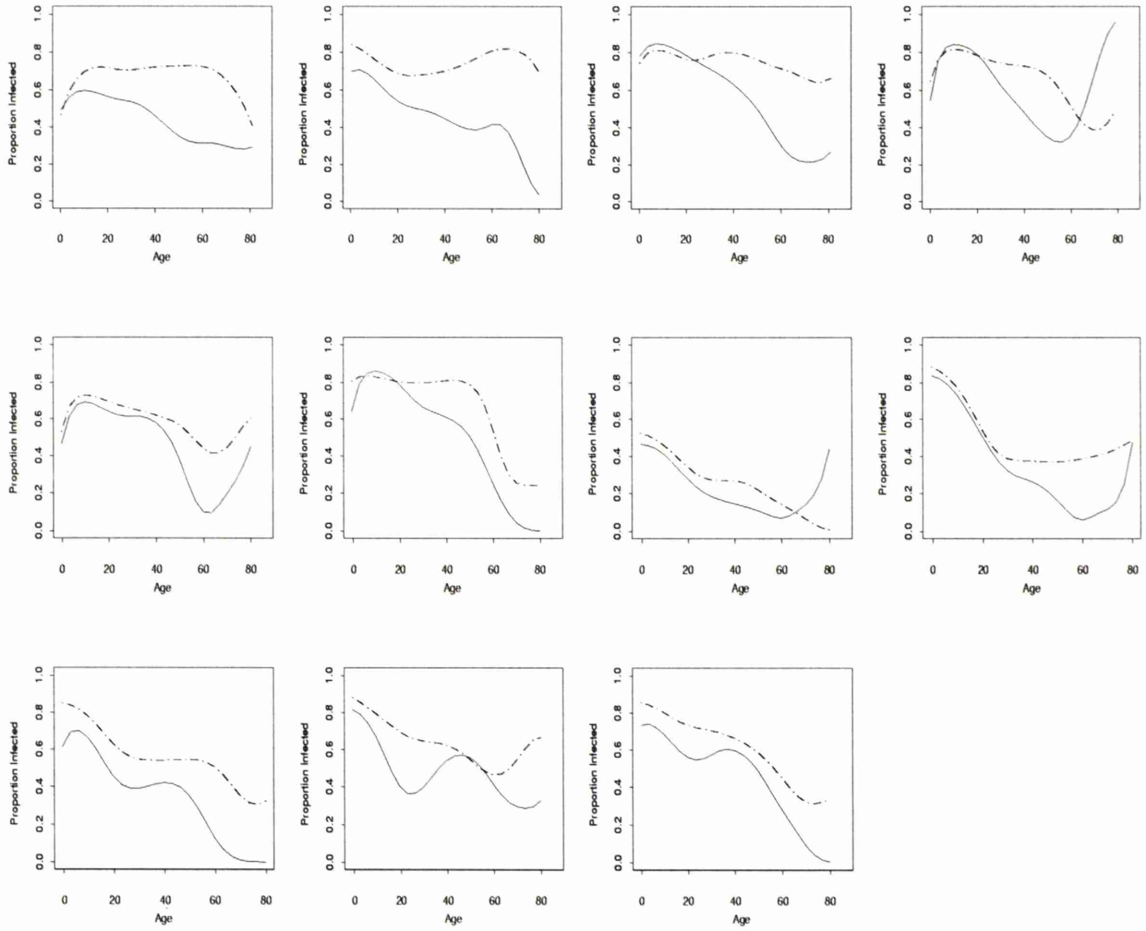
We fit a generalized additive model to the binary response of infection with smooth term for age and a linear effect for the categorical variable sex. Our fitting mechanism also takes the binomial variance into account. The fitted logistic models to Age for both villages Panchi and Laochi, along with the point wise 95% confidence bands are shown in the graph 5.1.

Figure 5.1: The additive logistic fit of prevalence to age for Panchi and Laochi. The dashed-curves are point wise 2 x standard-error bands



As one would expect, the highest risk of infection is for young children with a decreasing risk for adults over 40 years. This is confirmed from the graph 5.2 in which we can see the proportion infected for both sexes as well as from the non-parametric regression curves constructed in previous chapters.

Figure 5.2: Prevalence of ascariasis in Panchi, Continuous line: Males, dashed line: Females



It can be shown from the above graph that females are more exposed to the disease than males. Also the beneficial effect of the treatment at the period just immediately after treatment is evident.

Similarly to Panchi, the proportion infected in Laochi was higher for young children but lower for adults as it is shown graph 5.3.

It is apparent from this graph the beneficial effect of treatment in the three months following period especially in young children.

So, we fit an additive model to the data, whose output is shown in table 5.1. As we can clearly see from this table, there are two types of degrees of freedom defined as 'df' and 'Nonpardf' for the parametric and non-parametric contributions within terms respectively. The 'Nonpardf' term is an intuitively defined quantity and can take on fractional values. Our smoothing spline was requested to produce a fit with 4 df (the default) and it returned one with 2.8 and 3 for Panchi and Laochi respectively, which are quite close for our purposes. In addition, the column labeled 'NparChisq' represents a type of score test to evaluate the nonlinear contribution of the non-parametric terms. In this case, it seems to indicate that none of the nonlinear components are significant. Further details can be found in 'Statistical Models in S'(Hastie, Chambers).

Table 5.1: Non-parametric effect for age in Panchi and Laochi

Village	**Parameter**	**df**	**Nonpardf**	**NparChisq**	**P(Chisq)**
Panchi	s(Age)	1	2.8	5.13	0.14
Laochi	s(Age)	1	3	4.47	0.22

To confirm this result, we make use of a prognostic plot for assessing the fit of the linear model. That is, we draw the graph of the linear logistic model with its 95% confidence bands along with the non-linear fit(figure 5.4). As it is shown from graph 5.4, the fit of the non-linear model lies within the bands of the linear model, so we can presume that the latter is an adequate fit for our data. Hence, the model fitted for prevalence in Chapter

3,in which age is treated as a linear effect describes in a reasonable accurate way the data from this village. Similarly for Laochi, the non-linear effect of age is not significant which is shown in graph 5.5, where the linear model is displayed along with the non-linear one.

It is apparent from figure 5.5 that the linear model describes the data in a reasonable accurate way since the non-linear fit lies within the 95% point wise confidence bands of the linear model. Hence, we adopt the model with age as a linear effect as is fitted in chapter 3. At this point, we should notice that although the generalized additive models fitted to data from Panchi and Laochi check linearity assumptions more formally, they cannot be trusted as much as the generalized linear models fitted in Chapter 3 since the latter ones are based on a well-founded theory under the assumption of independent errors, while the former ones are approximate.

As for the Nigerian villages, graph 5.6 shows that there is a high risk of infection in school-age children.

Figure 5.3: Prevalence of ascariasis in Laochi

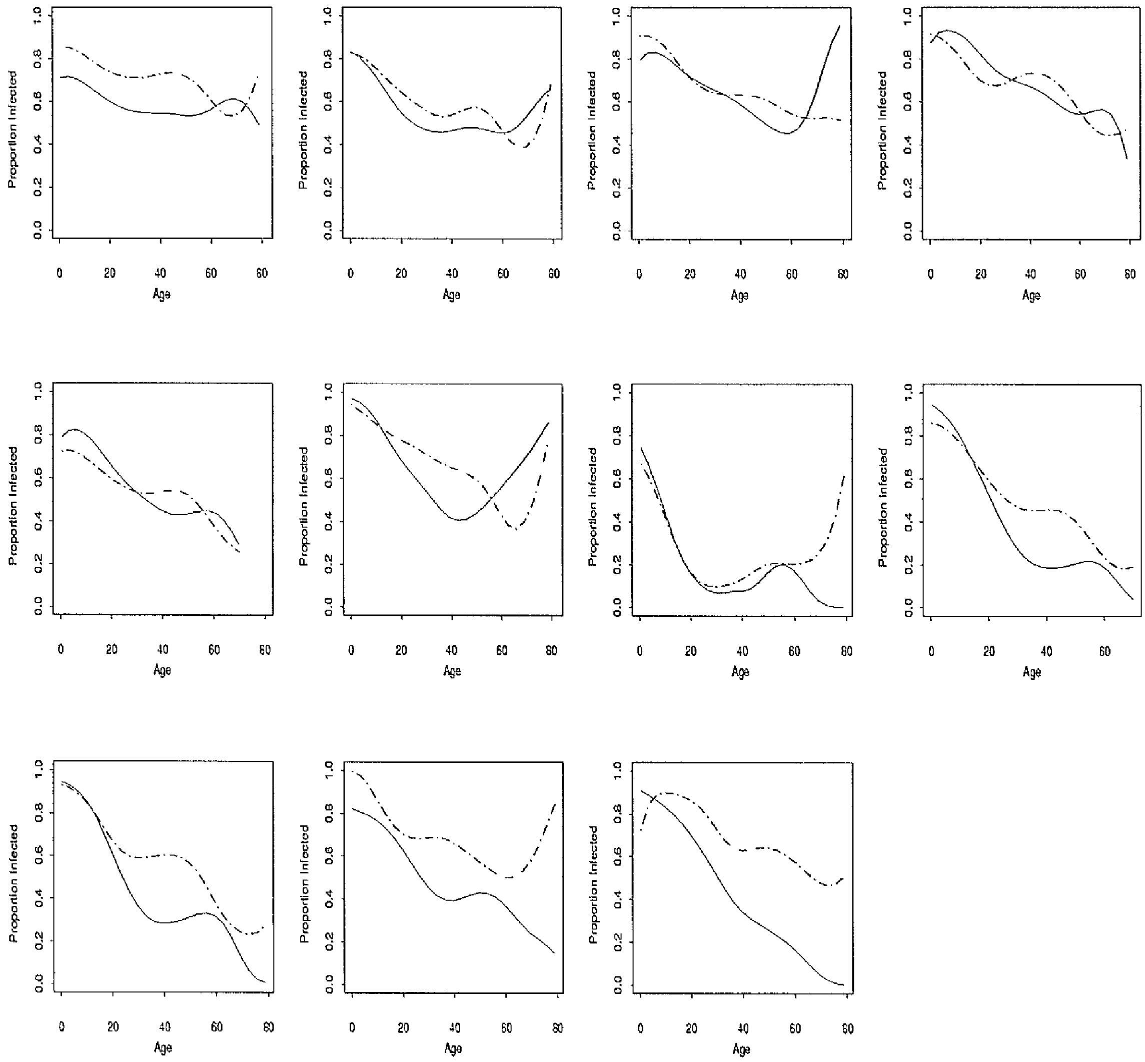


Figure 5.4: Comparison between linear and non-linear effect of age in Panchi for the period before treatment

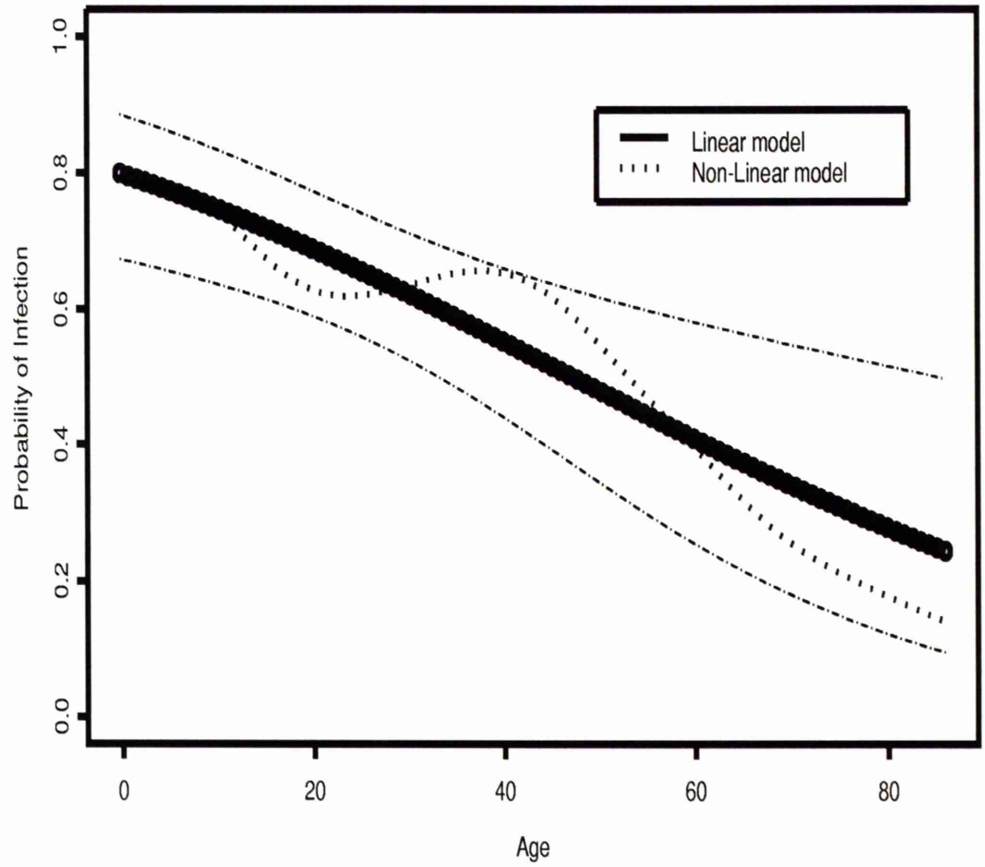


Figure 5.5: Comparison between linear and non-linear model in Laochi for the period before treatment

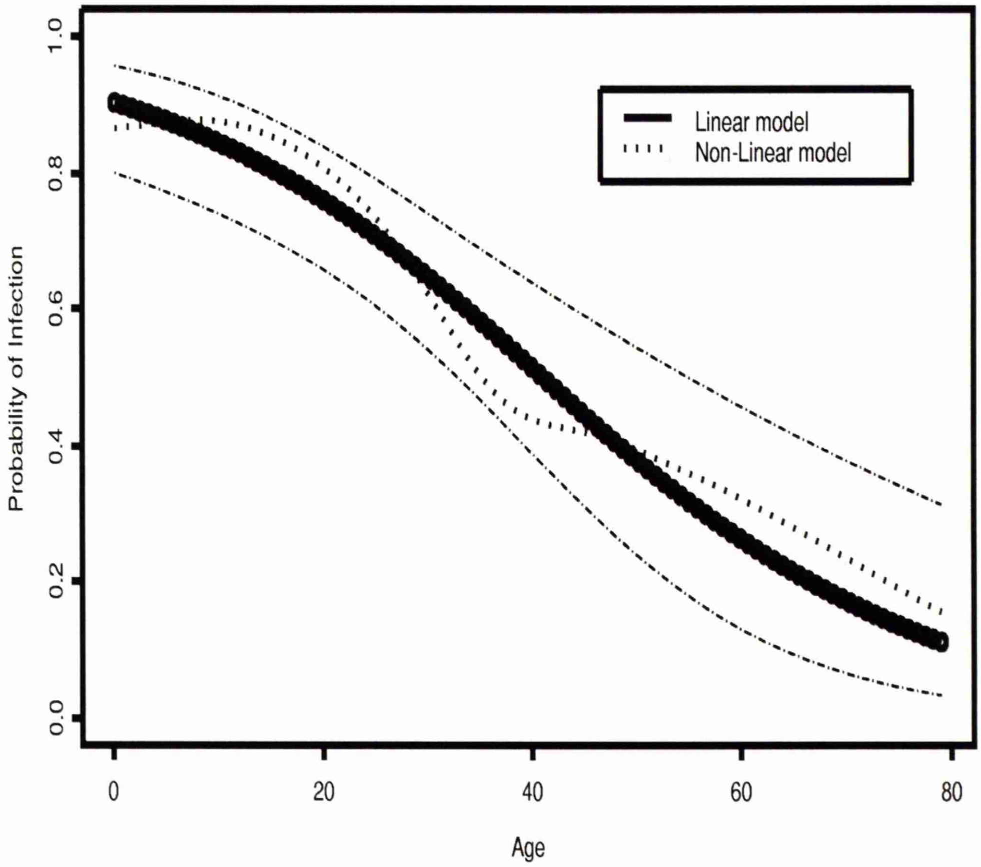
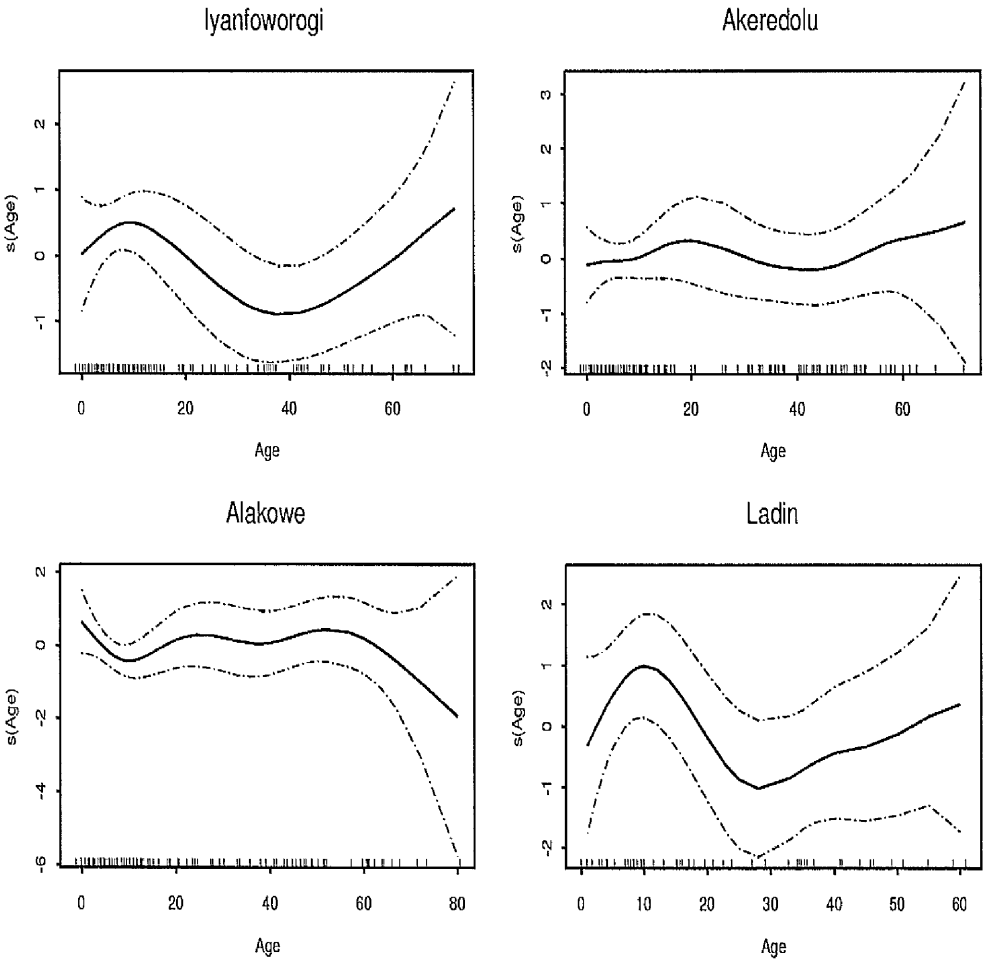


Figure 5.6: The additive logistic fit of prevalence to age for the Nigerian villages. The dashed curves are point wise 2 x standard error bands



It is apparent in the above graph that especially in Iyanfoworogi and Ladin the risk is higher for children aged 5 to 15 years but lower in adults. Table 5.2 shows the non-parametric contribution of age in all the four villages.

Table 5.2: Non-parametric effect for age in Nigeria villages

Village	**Parameter**	**df**	**Nonpardf**	**NparChisq**	**P(Chisq)**
Iyanfoworogi	s(Age)	1	2.9	8.13	0.04
Akeredolu	s(Age)	1	3	3.08	0.36
Alakowe	s(Age)	1	3	11.04	0.011
Ladin	s(Age)	1	2.8	9.03	0.025

As we can see from the above table, there are significant non-linear effects for age in Iyanfoworogi, Alakowe and Ladin. The next step will be to draw the fit of the linear model and its 95% confidence bands along with the non-linear fit in order to confirm the above results.

So, in Iyanfoworogi where targeted chemotherapy in children aged 5 to 15 years provided once a year, the effect of age is significantly non-linear as the following graph shows.

Figure 5.7: Linear and non-linear effect of age in Iyanfoworogi

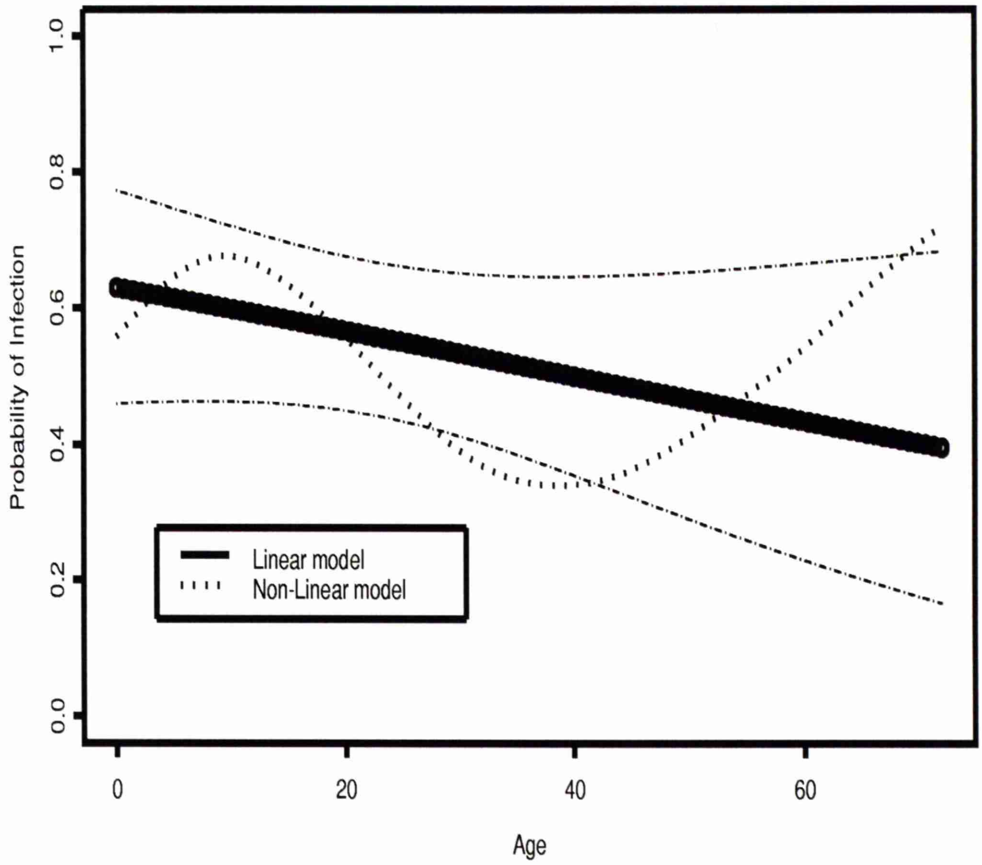
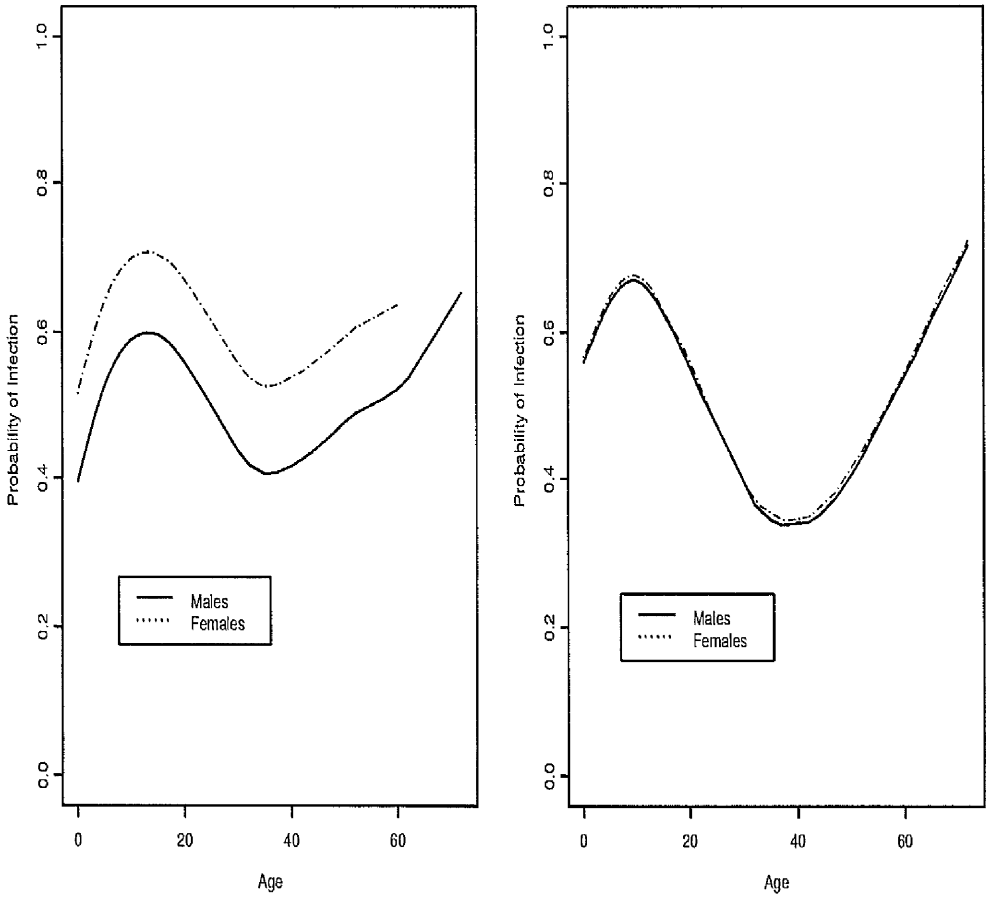
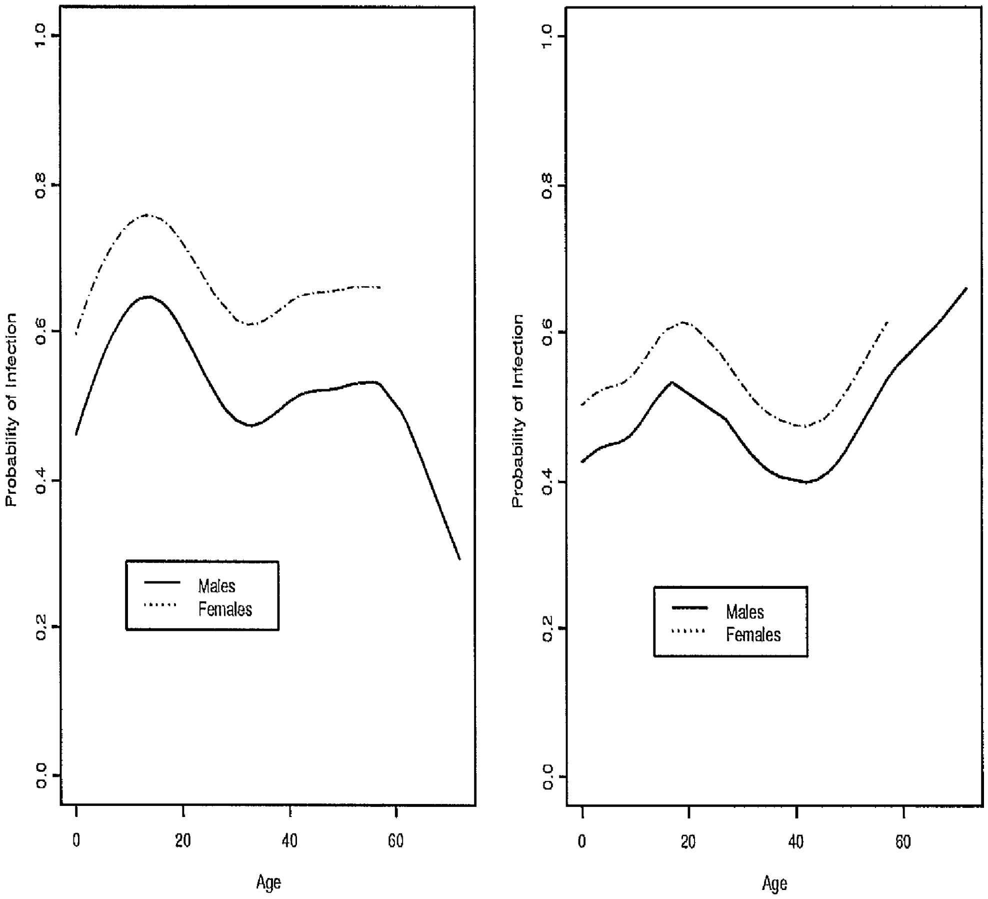


Figure 5.8: Prevalence of ascariasis in Iyanfoworogi before(left graph) and one year after treatment(right graph)



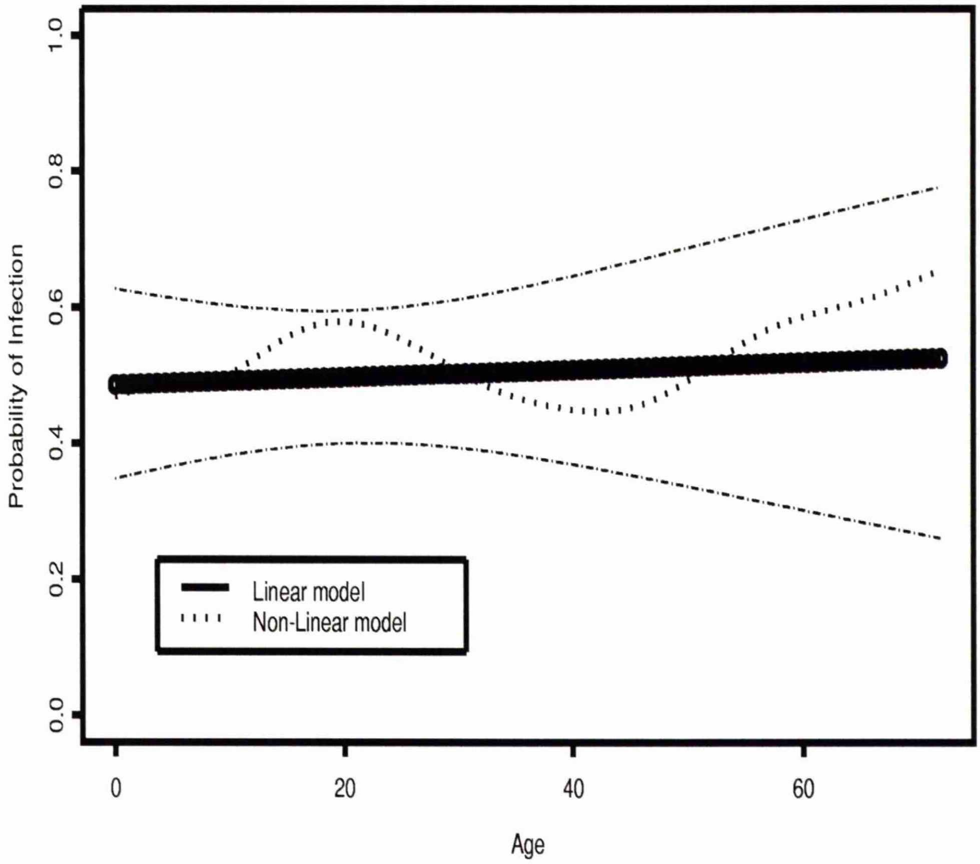
It is apparent from graph 5.7, that the linear model is in doubt since the non-linear fit lies outside the confidence bands of the linear model. Furthermore, the impact of the treatment does not seem to be beneficial as graph 5.8 shows.

Figure 5.9: Prevalence of ascariasis in Akeredolu



In fact, the infection rate for the targeted group is still high after treatment. Hence, the model fitted in Chapter 3 in which age is regarded as a factor with three levels does describe the data in a reasonable accurate way. On the other hand, in Akeredolu in which targeted chemotherapy was provided twice a year, the beneficial effect of treatment is evident, as the graph 5.9 shows.

Figure 5.10: Linear and non-linear effect of age in Akeredolu



As we can roughly see from graph 5.9, the prevalence of ascariasis in children was reduced significantly after treatment. Also, the effect of age is linear for this village, as it is shown in graph 5.10.

It is apparent from the graph 5.10 that the model in which age is treated as a linear effect describes our data accurately. Thus, another way of handling this dataset would be to fit a model with age as a linear effect as the following table shows.

Table 5.3: Estimated Parameters for Akeredolu

Variable	**Coefficient**	**Std. Error**	**p value**
Constant	0.518	0.41	0.1973
Age	0.0004	0.099	0.9682
Sex	-0.1363	0.39	0.7308

From Table 5.3, we can clearly see that the effects of both age and sex are not significant predictors for the prevalence of ascariasis. Hence, we would presume that although the impact of treatment was significant, as it is shown in figure 5.9, there were no significant differences among the age groups and also between males and females. Again, it should be mentioned that although the generalized additive model fitted to the data from this village suggests that age should be regarded as a linear effect, it cannot be trusted as much as the generalized linear model fitted for this dataset in Chapter 3, in which age is regarded as a factor. Unlike Akeredolu, the significant non-linear effect of age is evident in Alakowe and Ladin. Specifically in Alakowe where targeted chemotherapy was provided three times a year, there is a high risk in children, which is significantly non-linear as it is shown in graph 5.11 in which the non-linear fit exceeds the confidence bands of the linear model, while the impact of treatment in school-age children is evident. In this group the infection rate reduces significantly as we can clearly see in graph 5.12.

Figure 5.11: Linear and non-linear model for Alakowe

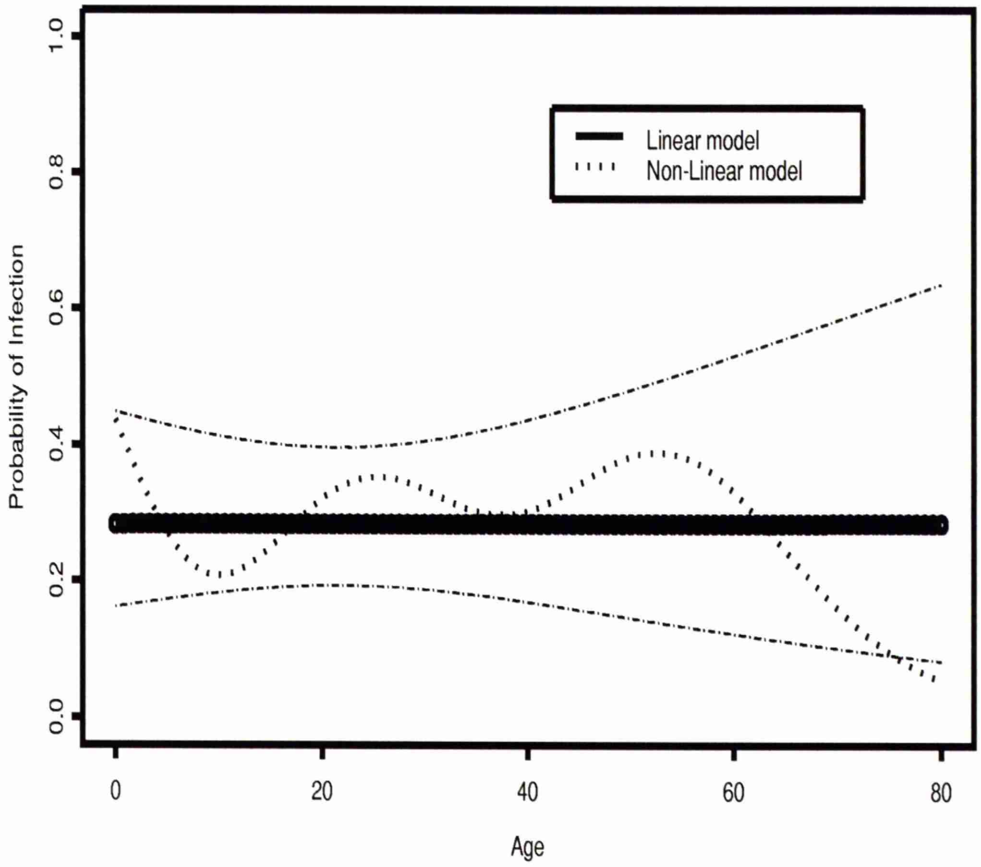
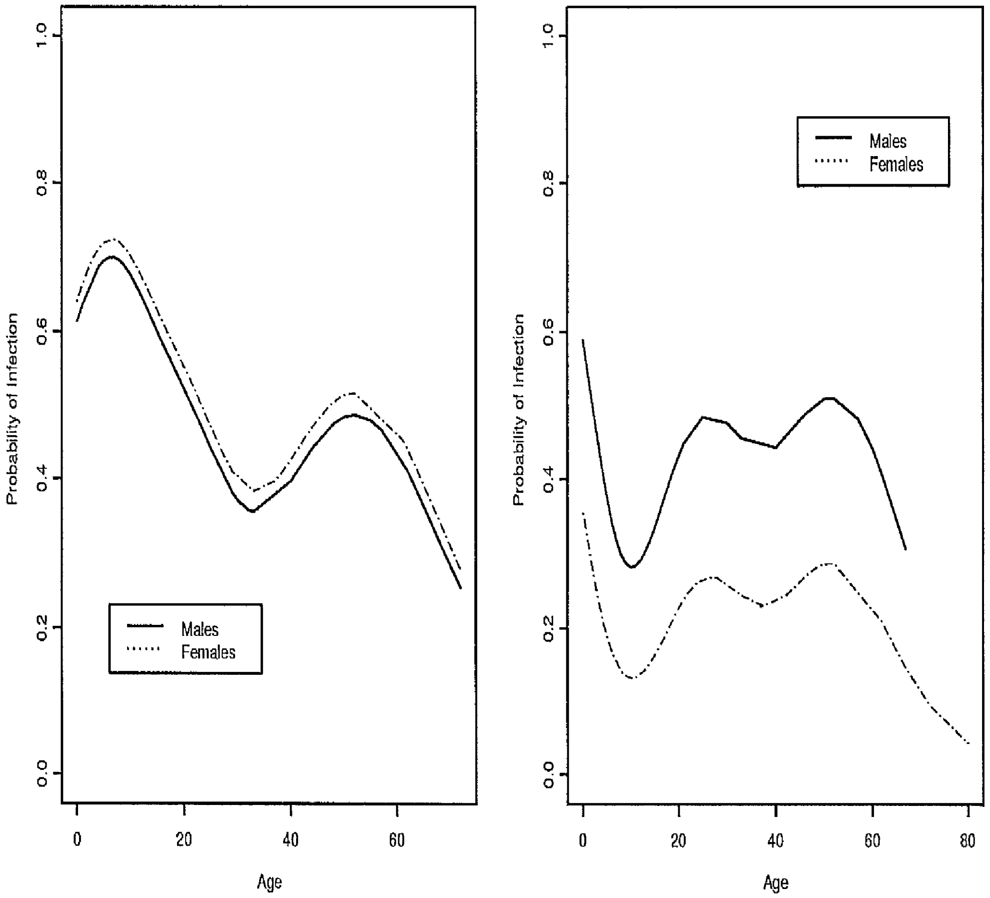


Figure 5.12: Prevalence of ascariasis in Alakowe



Finally in Ladin which was acted as a control village, we have similar results. There is a significant non-linear age effect which is apparent in graph 5.13 and also the prevalence of the disease does not decrease but it increases in children of both sexes after treatment (graph 5.14).Hence, the fitted models for Alakowe and Ladin, as have been described in Chapter 3, where age is regarded as a factor with 3 levels, describe the data in a reasonably accurate way.

Figure 5.13: Linear and non-linear model for Ladin

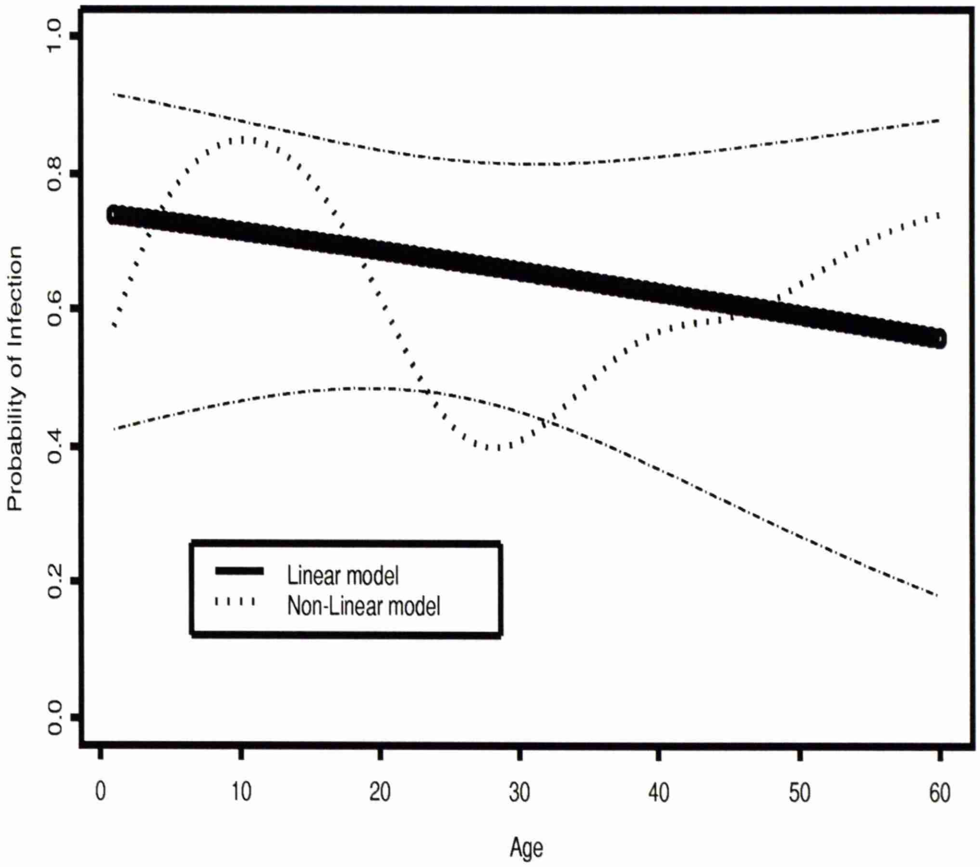
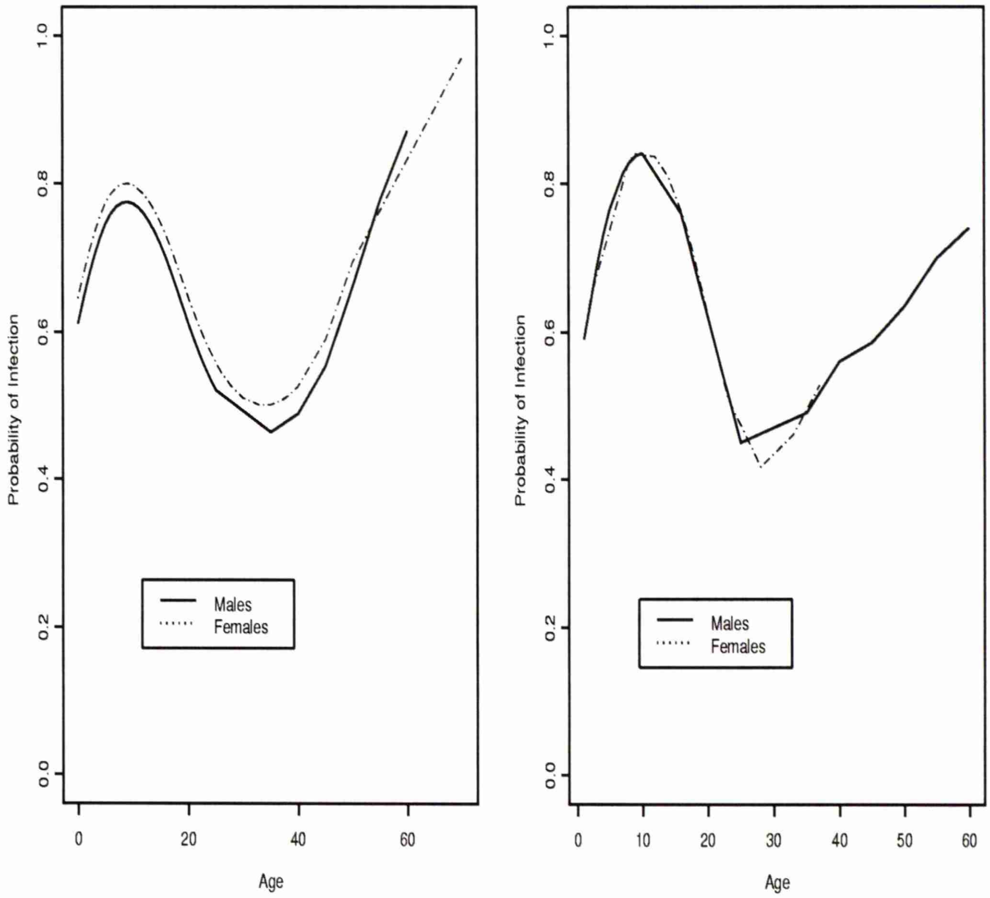


Figure 5.14: Prevalence of ascariasis in Ladin

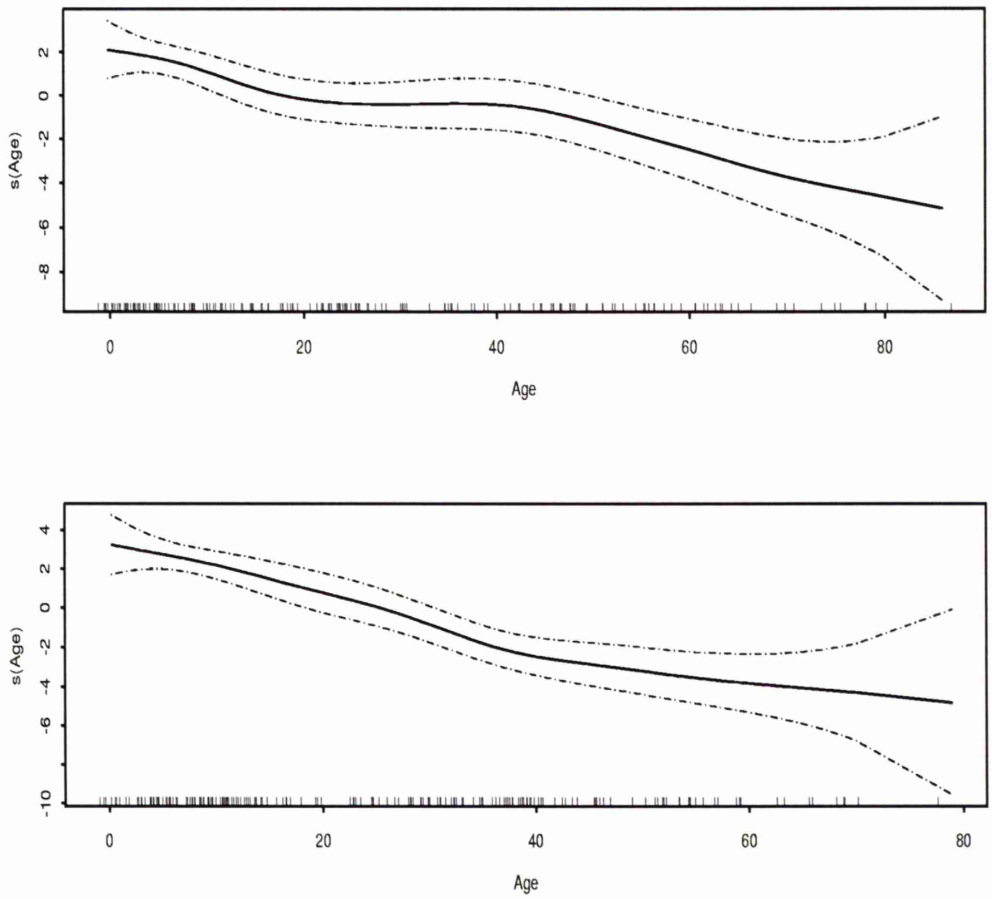


The above results concerning the prevalence of ascariasis in China and Nigeria, confirm our conclusions from previous chapters. Specifically in chapter 4 in which the size of the effect of the treatment was the main focus, we had also presumed that the infection rate drops significantly not only in China for children aged 5 to 15 years but also in Alakowe village of Nigeria. Also, the significant linear effect for age in Panchi, Laochi and Akeredolu justifies that the linear models fitted for these villages are adequate. On the other hand, there is a significant evidence for non-linear effect of age for data from Iyanfoworogi, Alakowe and Ladin. These effects can be assumed to be smooth, but the corresponding inferences are approximate therefore should not be trusted as much as the ones from usual linear and generalized linear models.

Intensity

Similar to prevalence, we fit an additive model to the log of the intensity to succeed stable variance with smooth term for age and a linear effect for sex . This time, our fitting mechanism takes normal errors into account. The smooth term is computed in the same way as above, that is by using a smoothing spline. The smoothing parameter is also set to default, which in this case is $df = 4$. As we can see in figure 5.15, the intensity of ascariasis in Chinese villages is higher in young children but lower in adults.

Figure 5.15: Fitted additive models of intensity for age in Panchi and Laochi. The dashed-curves represent point wise 2 x standard error bands



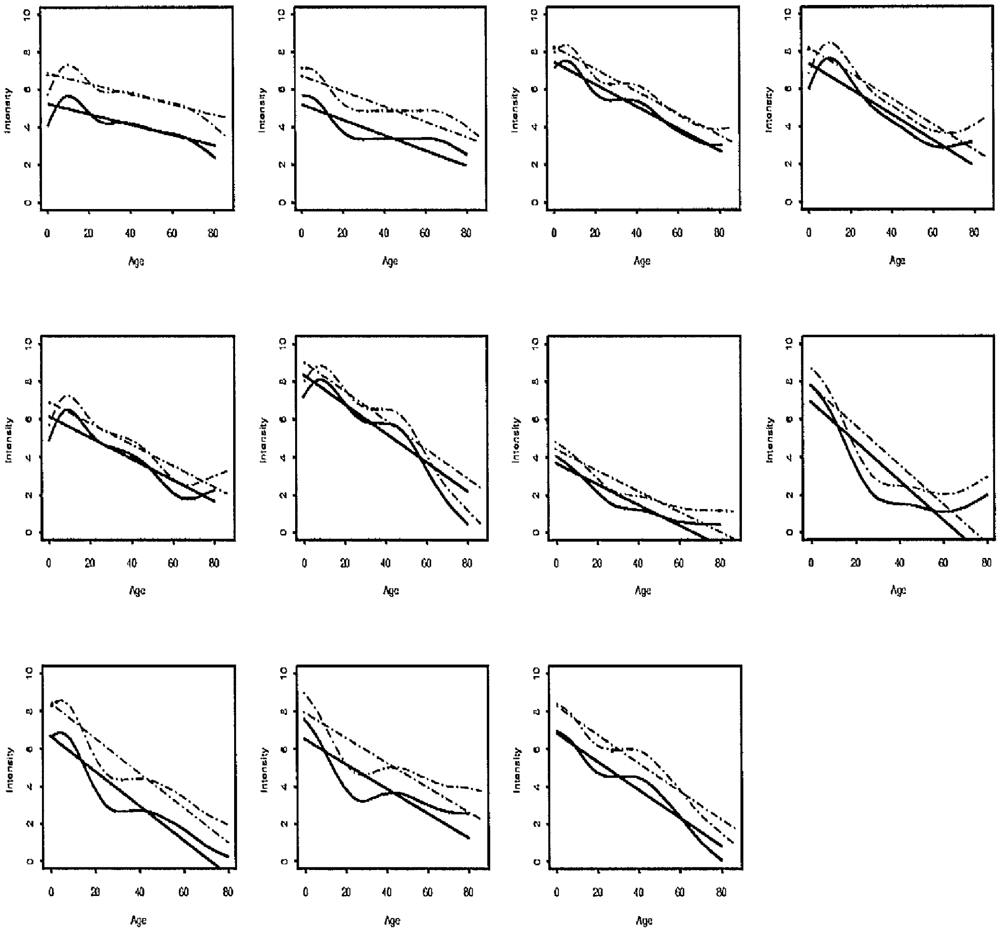
The output of the fitted models for the Chinese villages is displayed in table 5.4. As we can see from this table, the smoothing spline returned a fit with 3 degrees of freedom for the non-parametric contribution for age. The column 'NparChisq' shows that none of the nonlinear effects are significant.

Table 5.4: Non-parametric effect for age in Panchi and Laochi

Village	**Parameter**	**df**	**Nonpardf**	**NparChisq**	**P(Chisq)**
Panchi	s(Age)	1	3	1.67	0.17
Laochi	s(Age)	1	3	1.24	0.29

Therefore, the model, in which age is treated as a linear effect, describes accurately the data from both villages, as it is shown in the following graphs.

Figure 5.16: Intensity of acsariasis in Panchi



It is apparent from graphs 5.16 and 5.17 the beneficial effect of treatment at the period just immediately after treatment. The fitted model for the intensity in Panchi, where age is treated as a linear effect, shows that there is no significant difference between the age groups for the period just immediately after treatment(i.e. 7th survey), as Table 5.5 shows.

Table 5.5: Estimated Parameters for intensity in Panchi

Variable	Coefficient	Std. Error	p value
Constant	2.27	0.7	0.02
Age	-0.006	0.007	0.388
Sex	0.451	0.212	0.04

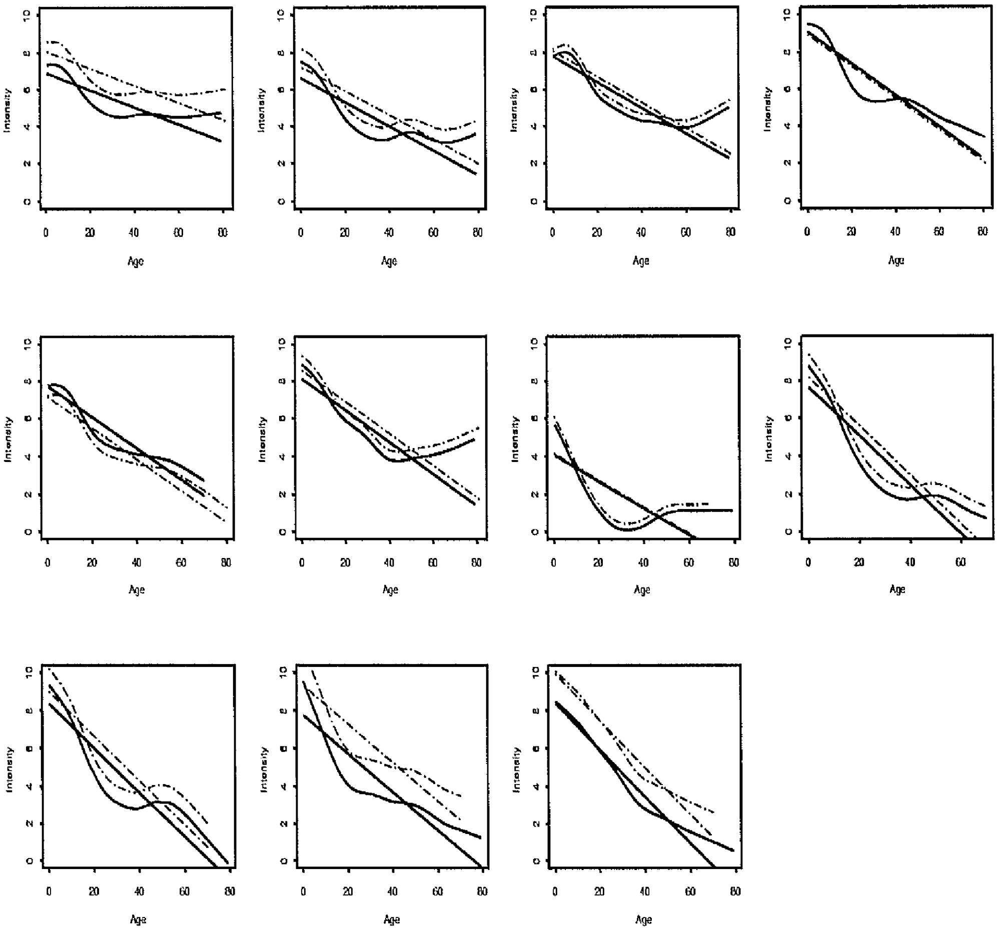
Unlike Panchi, the corresponding model in Laochi shows that there is a significant difference between the age groups for the post-treatment period.

Table 5.6: Estimated Parameters for intensity in Laochi

Variable	Coefficient	Std. Error	p value
Constant	3.22	1.09	0.006
Age	-0.02	0.01	0.031
Sex	0.46	0.294	0.127

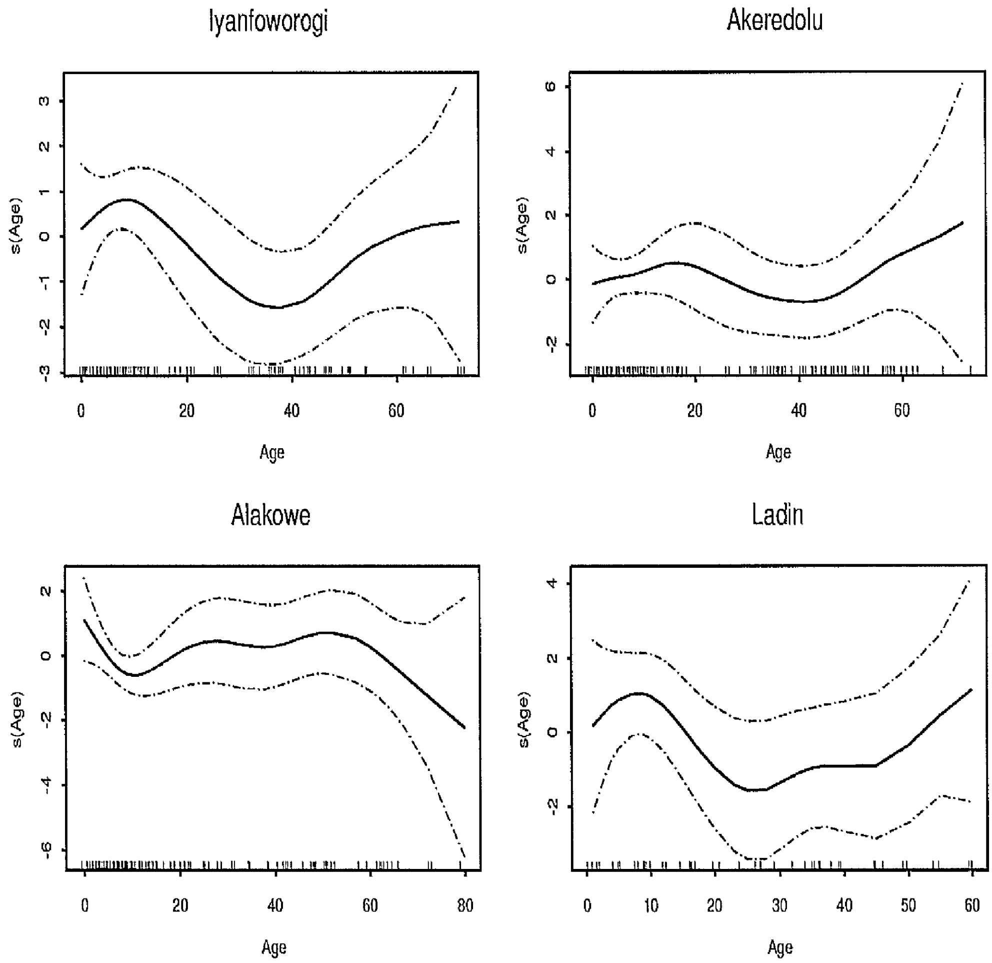
Similar to prevalence, the generalized additive models fitted for intensity, which suggest that age should be treated as a linear effect, cannot be trusted as much as the linear models due to the fact that in former ones any inference is approximate.

Figure 5.17: Intensity of ascariasis in Laochi



Similarly for Nigerian villages, we fit an additive model to the logarithm of the response from the gaussian family with a linear effect for sex and a non-linear effect for age. As we can see in the graph below, there is a higher risk in intensity for young children but lower for adults in all villages.

Figure 5.18: The additive fit of the log of intensity for age in Nigerian villages. The dashed-curves represent the point wise 2 x standard error bands



There is an evidence that the age effect is significantly non-linear for Iyanfoworogi and Alakowe but linear for the other two villages. These results are shown in table 5.7.

Table 5.7: Non-parametric effect for age in Nigerian villages

Village	**Parameter**	**df**	**Nonpardf**	**NparChisq**	**P(Chisq)**
Iyanfoworogi	s(Age)	1	3	2.72	0.047
Akeredolu	s(Age)	1	3	1.47	0.22
Alakowe	s(Age)	1	3	4.44	0.005
Ladin	s(Age)	1	3	2.6	0.063

So, in Iyanfoworogi where targeted chemotherapy was provided once a year, a model, as it has been fitted in Chapter 3, with age as a factor would describe the data in a reasonable accurate way. In the following graph, the intensity of ascariasis in this village before (left) and after treatment (right) is shown.

Figure 5.19: Intensity of ascariasis in Iyanfoworogi

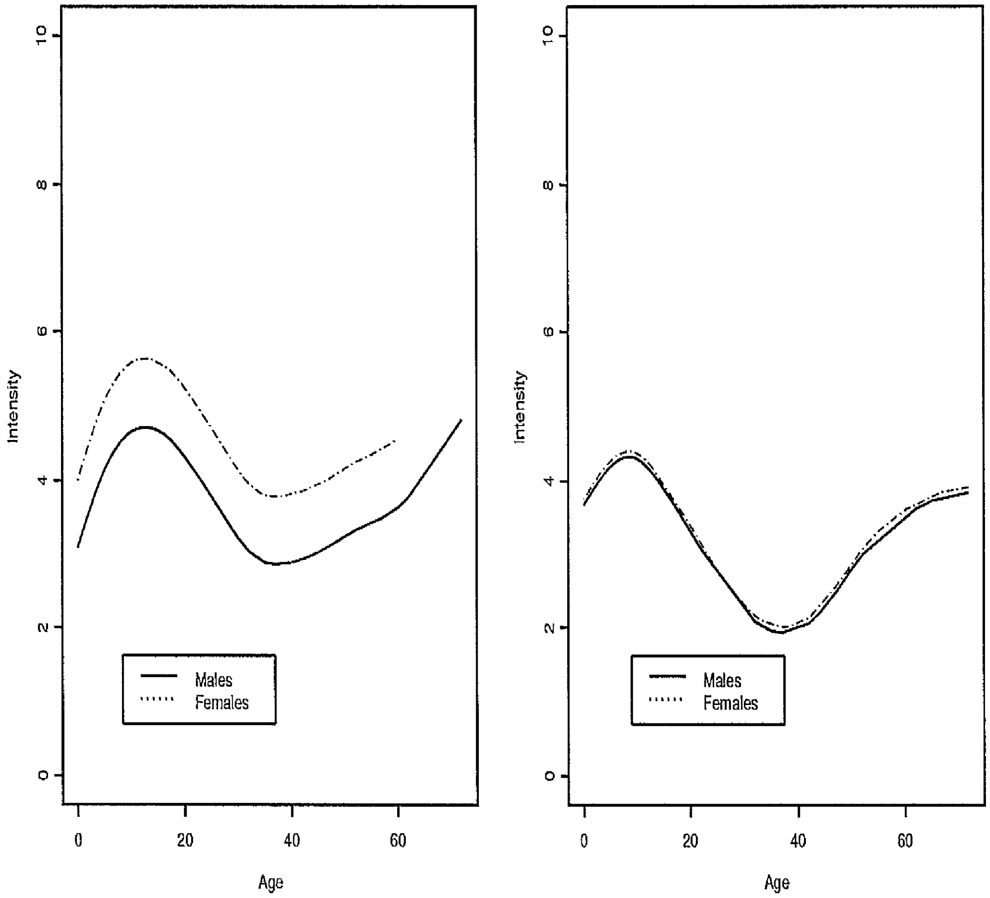
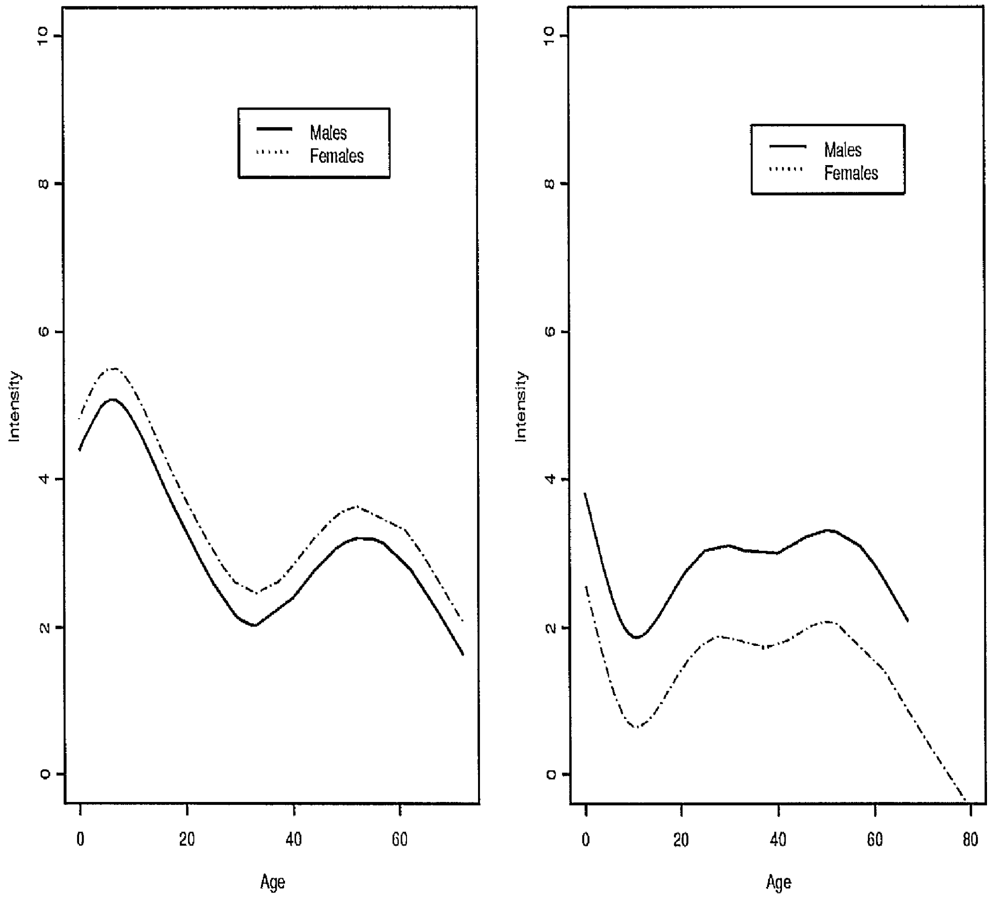
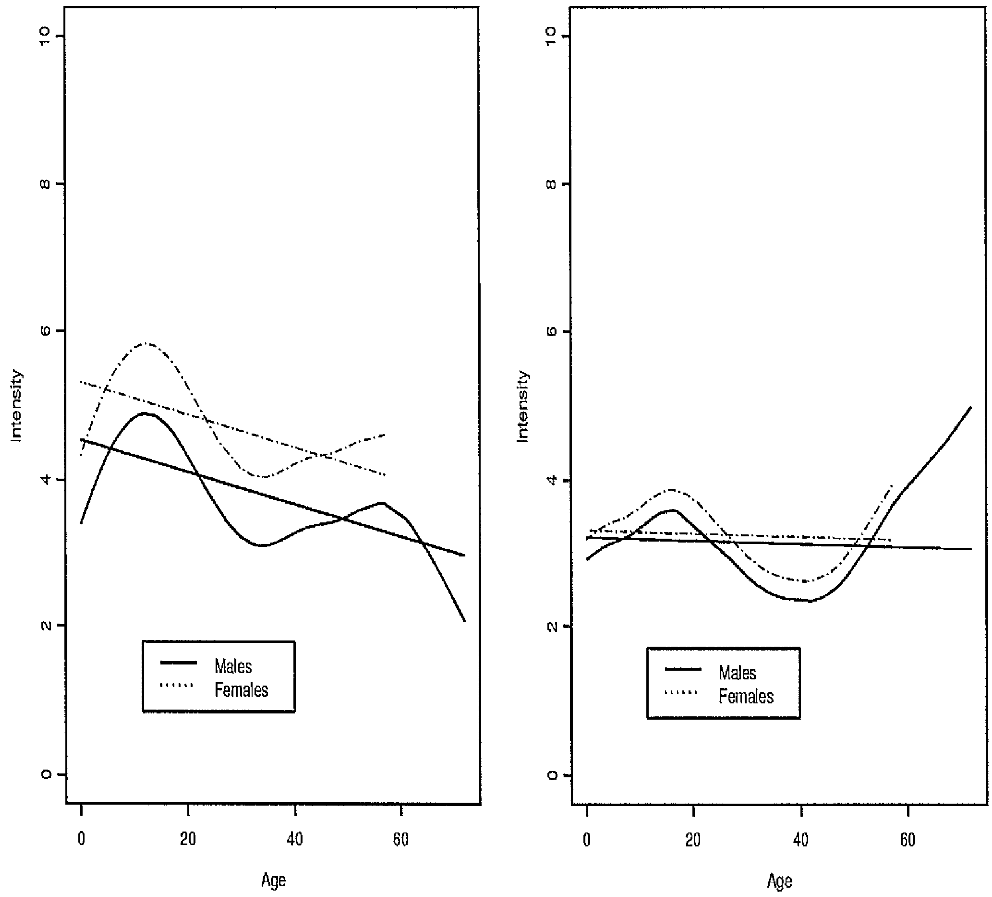


Figure 5.20: Intensity of ascariasis in Alakowe



It is apparent from graph 5.19 that similar to prevalence, the intensity of ascariasis does not drop significantly in children. There is also no significant difference between males and females for the period after treatment. Similarly, we fit an additive model with a non-linear effect for age in Alakowe in which targeted chemotherapy was provided three times a year. Graph 5.20 shows the intensity before and after treatment in this village.

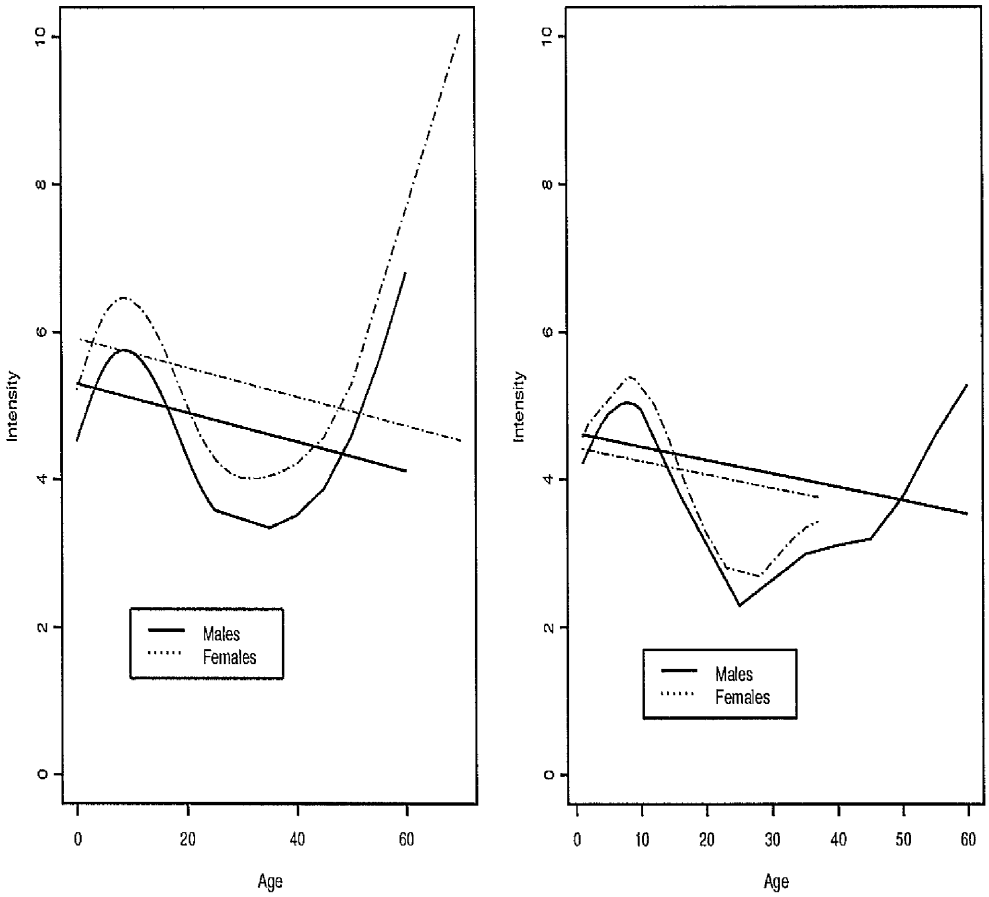
Figure 5.21: Intensity of ascariasis in Akeredolu



As we can see from graph 5.20, the intensity in children drops significantly after treatment. Similar to Iyanfoworogi, the model in which age is regarded as a factor would describe the data accurately.

On the other hand, the effect of age in the other two villages Akeredolu and Ladin is not significantly non-linear which means that in this case, the model, in which age is treated as a linear effect could be an adequate fit to the data, as it is shown in the graphs 5.21 and 5.22.

Figure 5.22: Intensity of ascariasis in Ladin



It is apparent from graph 5.21 that the effect of treatment does not differ among the age groups, as it is confirmed from the fitted linear model in chapter 3 (Table 3.20). There is also a reduction in intensity for the post-treatment period in Ladin, as the graph 5.22 shows, although no treatment was given in this village.

As we can see from the above graph, the linear model describes the data accurately. There is also no difference between males and females. The following tables show the estimated coefficients for the models which describe the intensity of the disease in these villages, where age is regarded as linear effect.

Table 5.8: Estimated Parameters for intensity in Akeredolu

Variable	**Coefficient**	**Std. Error**	**p value**
Constant	2.204	0.365	0.00
Age	0.0055	0.004	0.151
Sex	0.13	0.166	0.44

Table 5.9: Estimated Parameters for intensity in Ladin

Variable	**Coefficient**	**Std. Error**	**p value**
Constant	2.29	0.559	0.001
Age	0.007	0.005	0.179
Sex	0.223	0.163	0.189

As the tables 5.5 and 5.6 show, the effects of age and sex are not significant predictors for the intensity of ascariasis in Akeredolu and Ladin.

5.3 Conclusions

To sum up, we have seen that by using a generalized additive model either from the binomial or the gaussian family, not only can we draw conclusions about the effect of age in the prevalence and intensity of the disease, but we have also been able to confirm our assessments and conclusions made in previous chapters. Such a model is a useful descriptive tool and its usage is efficient for assessing the significance of a covariate, i.e in order to assess the linear or non-linear effect of age, we display the linear fit together with the non-linear one. It is also appropriate for the interpretation of a trend of a response variable, i.e by plotting the smooth function for age, we described the infection rate of ascariasis by age and sex. Finally, in cases where the linear model fitted for binary or continuous data is in doubt, then a generalized additive model would be an explanatory device to suggest a suitable class of transformations of covariates to be included in a generalized linear model. However, we cannot incorporate correlated errors as we did in Chapter 2 and also even when we assume independence of errors, the inferences drawn from such a models will be approximate.

Thus, among those models which describe the prevalence and intensity of ascariasis in Chinese villages, the ones with age as a linear predictor provide an adequate fit to these data sets. On the other hand, among these models which describe the prevalence and intensity of the disease in Nigeria villages, the ones with age as a linear predictor provide an adequate fit in some of the villages but not in all of them. So, for describing accurately the prevalence in Iyanfoworogi, Alakowe and Ladin, it would be better if we use a generalized additive model with time and sex as linear predictors but a smooth function for age as a non-linear covariate. Similarly for the intensity, an additive model would be more appropriate for describing the data in Iyanfoworogi and Alakowe with time and sex as linear effects and a non-linear effect for age.

To sum up, we could say that the conclusions drawn from this chapter are not very different from the ones in previous chapters. In fact, using generalized additive models we are able to confirm that the prevalence as well as the intensity of the disease are spread more in young children and less in adults and that the treatment is more effective in the former than in the latter age group in both males and females. Generalized additive models fitted in data sets from China and Nigeria are able to describe such trends in an appropriate way even when the data are non-linear. For instance, the effect of age is not always linear but it can be assumed to be smooth function without making any parametric assumptions. In this way, the results are similar with the ones obtained by the construction of non-parametric regression curves. However, one should really be careful when he or she interprets the results from these models, since generalized additive models are not based on a well-founded theory and any inference drawn from them will be approximate.

References

- Peng Weidong, Zhou Xianmin, D. W. T. Crompton(1995). Aspects of Ascariasis in China, Helminthologia, 32, p. 99, 1995

- Crompton(1991). The challenge of parasitic worms, pp 75-80

- C.V. HOLLAND, S.O. ASAOLU, D.W.T. CROMPTON, R.R. WHITEHEAD, I. COOMBS. Targeted anthelminthic treatment of school children : effect of frequency application on the intensity of A.lumbricoides infection in children from rural Nigeria villages.

- ASAOLU, HOLLAND, CROMPTON. Community control of Ascaris in rural Oyo State, Nigeria: mass, targeted and selective treatment with levamisole, Parasitology 103, p. 291-8

- Schultz, M.G. Schultz(1985). Ascariasis and its public health significance, Biology of Ascaris, p. 10-20.

- Peng Weidong, Zhou Xianmin, Cui Xiaomin, D. W. T. Crompton, R. R. Whitehead, Xiong Jiangqin, Wu Haigeng, Peng Jiyuan, Yang Yang, Wu Weixing, Xu Kaiwu and Yan Yongxing. Ascaris, people and pigs in a rural community of Jiangxi Province, China, Parasitology (1996) 113, p.545-557.

- A. Azzalini (1994). Logistic regression for autocorrelated data with application to repeated measures, *Biometrika 81, 4, pp 767-75*

- A. Azzalini, M. Chiogna (1995). Some S-PLUS tools for the exploratory and parametric analysis of repeated measures data, pp 26-27.

- A. W. Bowman & A. Azzalini. Applied Smoothing Techniques for Data Analysis, *Oxford Statistical Science Series, 18, pp 117-23*

- Harper (1998). A tutorial example on applying the logistic model to cohort studies. *Analytic Epidemiology - Part II, pp 18-20*

- B. S. Everitt. The analysis of repeated measures: a practical review with examples, pp 117-129

- Little & Rubin. Likelihood approaches to the analysis of missing data, pp 84-87, 129-130.

- G. W. Snedecor, W. G. Cohran(1980). Statistical methods: The comparison of proportions in paired samples.

- P. McCullagh and J.A. Nelder. Generalized Linear Models, pp 70-71, 110-111.

- P. J. Diggle, Kung-Yee Liang and Scott L. Zegger(1994). Analysis of longitudinal data, pp 42-43, 107-108, 111,131-136.

- T. Hastie & R. Tibshirani (1990). Generalized additive models, pp 97, 98, 140.

- J. M. Chambers, T. J. Hastie (1992). Statistical models in S, pp 252-306.