

UNIVERSITY OF NAIROBI

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES

UNIVERSITY OF NAIROBI
CHIROMO LIBRARY

SCHOOL OF MATHEMATICS

University of NAIROBI Library



0506281 5

LONGITUDINAL DATA ANALYSIS AND ITS APPLICATION IN EVALUATION OF WORMS CONTROL IN CHILDREN

A PROJECT SUBMITTED TO THE SCHOOL OF MATHEMATICS IN PARTIAL FULFILLMENT
FOR A DEGREE OF MASTER OF SCIENCE IN MATHEMATICAL STATISTICS

©JULY 2007

Declaration

I, the undersigned, declare that this project is my original work and to the best of my knowledge has not been presented for the award of a degree in any other University.

MURIUNGI ROBERT GITUNGA

Reg. No I56/7317/2005

[Signature]

23/08/07

Signature

Date

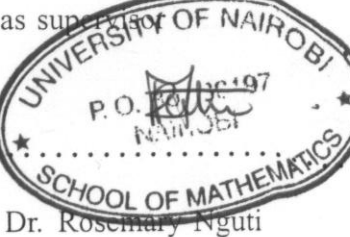
Declaration By Supervisor

This project has been submitted for examination with my approval as supervisor

Signature

Dr. M. Gacii
School of Mathematics
University of Nairobi

Signature



Dr. Rosemary Nguti
School of Mathematics
University of Nairobi

Acknowledgments.

I am very grateful to all those who have contributed in one way or the other towards the success of this work. Here are just but a few ; Dr. M. Gacii for allowing me to do the work under his supervision. Thanks goes to him for starting me up with this work and advising me on areas to get the data. Dr. Rosemary Nguti for her assistance and her role as a supervisor in the absence of Dr. M. Gacii. My deepest gratitude for her advice, guidance and patience throughout the study and reading through the work, Members of staff in the school of Mathematics for their guidance and constructive feedback during the presentations. My classmates for helping me out at one time or the other.

Dedicated to:

To my wife Lilian and the two little angels, Sharon and Brian, for peace and support they have offered

throughout the study.

1 GENERAL INTRODUCTION

2

1.1 Introduction	2
------------------	---

1.2 Intervening study	4
-----------------------	---

1.3 Longitudinal data	4
-----------------------	---

1.4 Fixed occasions	2
---------------------	---

1.5 Objectives	5
----------------	---

1.6 Significance of the study	6
-------------------------------	---

2 LITERATURE REVIEW

7

3 METHODOLOGY

9

3.1 Introduction	9
------------------	---

3.2 Random intercepts and slopes	10
----------------------------------	----

3.3 Estimation of parameters for random intercepts and slope model	12
--	----

3.4 Within subject and between subject variables	14
--	----

3.5 The logistic model	18
------------------------	----

3.6 Estimation of parameters for logistic model	20
---	----

3.7 Probit models	22
-------------------	----

3.8 Maximum likelihood estimation for the probit model	24
--	----

Contents

1	GENERAL INTRODUCTION	2
1.1	Introduction	2
1.2	Motivating study	4
1.3	Longitudinal data	4
1.4	Fixed occasions	5
1.5	Objectives	5
1.6	Significance of the study	6
2	LITERATURE REVIEW	7
3	METHODOLOGY	9
3.1	Introduction	9
3.2	Random intercepts and slopes	10
3.3	Estimation of parameters for Random intercepts and slope model	12
3.4	Within subject and between subject variables	14
3.5	The logistic model	18
3.6	Estimation of parameters for logistic model	20
3.7	Probit models	22
3.8	Maximum likelihood estimation for the probit model	24

4	DATA ANALYSIS	27
5	CONCLUSION	32
5.1	Recommendation for further study	33
5.2	REFERENCES	34

Abstract.

A logit and probit models were fitted to evaluate the control of worms in children.

The study considered four types of worms namely ascaris (as) hookworm (hw) schistosomula (sm) and trichuris (tr) which were observed in children for three times. Observations are based on dummy variables; presence of worms on t^{th} individual at t measurement occasions is denoted by 1 and 0 otherwise.

Four demographic factors on each individual were changed to be in binary form.

- i individual past treatment to worms (treat) i.e. whether a child has ever been treated to worms prior to the study or not. treated (1) and (0) otherwise
- ii pupils level in school i.e class five (0) or six (1).
- iii sex; male (1) or female (0).
- iv supplement (supl) i.e. whether the child was given the macronutrient or not; given (1) not given (0). The result was tabulated. An appropriate analytical models were described and fitted .The data was analyzed using SAS. The result was discussed and the conclusions made.

UNIVERSITY OF NAIROBI
CHIROMO LIBRARY

Chapter 1

GENERAL INTRODUCTION

1.1 Introduction

Bundy D.A.P. (1989) argued that one quarter of school going children are infected with worms which in turn affect their cognitive ability. This problem is high in poor countries.

Its with this context that we base our study on worms infection in children in Kenya and focus our attention in Nyanza province specifically Bondo district where poverty level is comparatively high.

In scientific context, a prediction is a rigorous often (quantitative) statement forecasting what will happen under specific conditions typically expressed in the form if A is true, then B will also be true. The scientific method (i.e. body of techniques for investigating phenomena and acquiring new knowledge as well as for correcting and integrating previous knowledge) which is based on gathering observable, empirical and measurable evidence subject to specific principles of reasoning, is built on testing assertions which are logical consequences of scientific theories.

This is done through repeatable experiments or observational studies. Mathematical models therefore, are frequently used to both describe the behaviour of something, and predict its future behaviour.

In this study, chapter 1 describes various terms and gives the objectives and significance of the study. Chapter 2 describes various studies that have been carried on worms in children giving the methods studies have adopted to analyze the data. Chapter 3 gives the various methods used to such data and describing the appropriate models for analyzing longitudinal data where the outcome response variables are binary and describes the appropriate method to be applied in estimation of regression parameters.

Chapter 4 deals with the application of above appropriate models in evaluation of the data set (worms control data) giving the result. It deals with computational analysis i.e. use of SAS software in analysis of the data. Chapter 5 presents the conclusion.

1.2 Motivating study

A data of 117 class five and six pupils collected at random in Bondo district primary schools by Kenya Medical Research Institute (KEMRI) is used. Four types of worms namely ascaris (as) hookworm (hw) schistosomula (sm) and trichuris (tr) are to be observed three times ($t=0,1,2$) on each individual

At time $t=0$, children were observed, some were found to have specific worms while others not. Their background information was taken. These information includes age, pupils class level, gender, control information in terms of treatment and macronutrient supplement dosage.

Later after 4 months. $s(t)=4$, they were observed again, some were found to have the worms while others not. This was repeated again after another 4 months.

1.3 Longitudinal data

Correlated response data arise in longitudinal studies where individuals are observed on at least two different occasions.

Correlated data commonly arise, for instance in biomedical studies and economics. In addition to the production of cross-sectional estimates, these data may be used, for instance in event history modeling, which may be used to uncover determinants of survival for individuals afflicted with a serious health condition or to estimate gross flows (important in the study of labor market dynamics)

1.4 Fixed occasions

Designs involving fixed occasions have fixed set $t=1,2,...,m$ of measurement occasions. This could be study of some educational or therapeutic programme; intake test, pre-test, mid programme test and follow-up test ($m=5$) or study of altitude change in early adulthood where altitude is measured shortly after each birthday from 18th to 25th year of age ($m=8$).

If data are complete, each individual provides information on all these occasions. However, sometimes data are incomplete and its assumed such absent data are missing at random.

chapter 3 will give some of the commonly used models for binary data.

1.5 Objectives

1. The primary objective of this study is to use an appropriate design to analyze a repeated measurement where the measured occasion is fixed.
2. The study involves identifying factors which affect worms infestation in children.
3. The study models the association between the dependent outcome responses i.e whether there is any relationship on different types of worms infections.

1.6 Significance of the study

Investing in children's health is a sound economic decision for governments to take, even if the moral justifications for such programmes are not considered. This will result to better educated and more productive adults. Therefore poor health can be confirmed to be a mechanism for the intergenerational transmission of poverty.

With this regard, worms infections being one of the major threat to school going children requires an attention by statisticians.

Chapter 2

LITERATURE REVIEW

Nokes C., Grantham M.S., Sanyer A.W., Cooper E.S. and Bundy D.A.P. on their paper "Parasitic helminth infection and cognitive function in school children" (1992) examined the effect of moderate to high worms burden of *Trichuris trichiura* infections on cognitive functions of 159 school children aged 9-12 years in Jamaica using a double-blind placebo controlled protocol.

They evaluated results using a forward-stepwise multiple linear regression and concluded that removal of worms led to a significant improvement in tests of auditory short term memory. Nine week after treatment , it was concluded that there was no longer significant differences between the treated children and infected control group.

Lon K. Kightlinger, John R. Seed and Mynna B. Kightlinger (1995) studied *Ascaris lumbricoides* intensity in relation to environmental, socio-economic and behavioral determinants of exposure to infection in children from south East Madagascar.

They used *ascaris lumbricoides* counts as the outcome products of exposure proxy variables on 4-10 years age old children living in south East Madagascar.

The study reveals prevalence of 93% for *ascaris lumbricoides* 55% for *trichuris* and 27% for hookworms. They also carried a concurrent socio-economic household survey by visitation

and interview. Exposure to infection was assessed by environmental, demographic, behavioral and socio-economic indicators. This revealed that *ascaris lumbricoides* aggregations were associated with gender, housing styles, ethnicity and agricultural factors. Therefore they suggested that exposure and infections are obliquitors in their child population and that *Ascaris lumbricoides* intensity is influenced by gender-related behavioral and environmental factors that contribute to exposure.

They carried multiple regression where worms counts were transformed to common logarithms. they used SAS for computational and analysis.

2.1 Introduction

There are several different approaches available for implementing a repeated measures approach that models discrete data. Repeated measures designs are frequently used in medical research to account for the correlation in the data. They are often sets of draw back. When the response variable is binary, least square models are used to analyze longitudinal data more so when the response variable is binary. This is because the response variable in this case is restricted to lie between the values zero and unity. Least square estimation of the parameters may lead to fitted values of the response outside this range which would be clearly unsatisfactory. Another problem is the assumption of error terms with equal variance which is unlikely to hold in repeated measures data.

Due to such problems, it is more useful to postulate a model for the response variable after applying a transformation.

Chapter 3

METHODOLOGY

3.1 Introduction

There are several different approaches available for implementing a repeated measures approach that models discrete data. Repeated measures designs are frequently used in medical investigations to account for the correlation in the data. There are often sets of draw backs when ordinary linear models are used to analyze longitudinal correlated data more so when the response variable is binary. This is because, the response variable in this case is constrained to lie between the values zero and unity. Least square estimation of the parameters may lead to fitted values of the response outside this range which would be clearly unsatisfactory. Another problem is the assumption of error terms with equal variance which is unlikely for this type of variance.

Due to such problems, its more useful to postulate a linear model for the response variable after applying a transformation.

The two most appropriate methods is where we take the logit function which transforms the linear predictor to lie between 0 to 1 or we use the normal distribution to be the tolerance distribution in probit model. In such a case therefore, parameter estimates may now be obtained by maximum likelihood methods using generalized estimating equation (GEE). In the following subsections, we will discuss the application of some of these methods in the context of our study and then seek to justify the approach that we finally adopted.

In analysis of repeated measures, the commonly used models are the Random intercepts and slopes, Within subject and between subject variables, logistic and probit models. These models are discussed in the following subsections, describing their estimation procedures.

3.2 Random intercepts and slopes

Guldmond(1994) used a model with 2 parts i.e. random intercepts and a slope to analyze a repeated measures data.

He described the unexplained variability in single level multiple regression analysis as just the variance of the residual terms. The model take the form

$$Y_i = \beta_{0j} + \beta_{ij}.x_{ij} + R_{ij} \quad (3.1)$$

where Y_i is the outcome response of individual i , β_{0j} are intercepts and β_{ij} are regression coefficients or slopes.

If we wish to reduce the unexplained variability, we expand the regression equation (3.1) by predicting the group dependent regression coefficients β_{0j} and β_{ij} from level 2 variables z where z represent the explanatory variables or covariates such as school and if we have one such variable, the regression formulae for β_{0j} and β_{ij} on the variable z are given as

$$\beta_{0j} = \gamma_{00} + \gamma_{0i}.z_j + U_{0j} \quad (3.2)$$

$$\beta_{ij} = \gamma_{10} + \gamma_{11} \cdot z_j + U_{ij} \quad (3.3)$$

β_s here are treated as dependent variables in regression models for the population of groups.

(3.2) represents intercepts as outcome model.

(3.3) represents slope as outcome model.

substituting (3.2) and (3.3) in equation (3.1), we get

$$y_{ij} = \gamma_{00} + \gamma_{01}z_j + U_{0j} + (\gamma_{10} + \gamma_{11}z_j + U_{ij})x_{ij} + R_{ij} \quad (3.4)$$

which is equivalent to

$$y_{ij} = \gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + \gamma_{11}z_jx_{ij} + U_{0j} + U_{ij}x_{ij} + R_{ij} \quad (3.5)$$

However in multilevel data, variability has a more complicated structure because several populations are involved. Explaining variability therefore, can be achieved by explaining variability between individuals and variability between groups. The part of residual variance that is expressed by $\sigma^2 = \text{var}R_{ij}$ can be diminished by including other level 1 variables different from group to group. Inclusion of such variables may also diminish residual variance at group level. Another possibility is to try to find explanations in the population of group at level 2 which is represented by U_{ij} in equation (3.4).

3.3 Estimation of parameters for Random intercepts and slope model

Laid and Ware(1982), Goldstein(1986) and Longford(1987) have proposed a number of iterative estimation algorithms. Where the parameters of the random part together with σ^2 are known, then regression coefficients could be estimated straightforwardly with generalized least square (GLS)method.

Where all regression coefficients are known, total residuals could be computed and their covariance matrix could be used to estimate the parameters of the random part.

Use of iterated generalized least square method (IGLS) is more appropriate which involve use of provisional values for the random part parameters to estimate regression coefficients, then use the latter estimates to estimate the random part parameter again (and now better) then go on to estimate the regression coefficients again and again and so on until convergence of this iterative process. Other methods include the Fishers scoring method (Longford, 1993a), the Expectation-Maximization (EM) method (Bryk and Raudenbush, 1992) and the Generalized Estimating Equations method (Liang and Zeger, 1986). The latter is briefly described below.

The Generalized Estimating Equations method.

For a model based with discrete factors, the generalised estimating equation (GEE) method of parameter estimation is specifically more efficient for statistical hypothesis testing with correlated longitudinal data. GEE are parameter estimating equations set to account for the correlation between observations in generalised linear regression models and builds on variance estimation developed to protect against inappropriate assumptions about the variance. When data are collected on the same units across successive points in time, these repeated observations are correlated over time. If this correlation is not taken into account, then the statistical standard error of the parameter estimates will not be valid and hypothesis testing results will be non replicable.

In defining regression model using GEE methodology, one needs to define the following;

1. Distribution of the dependent variable. (Must be a member of exponential family)
2. Link function.
3. Independent variables.
4. Covariance structure of the repeated measurements.

3.4 Within subject and between subject variables

Maxwell and Delany(1990) used a model where they described the two types of variables.

Expected value for the measurement occasions t can be denoted by μ_t , and this model can be expressed by

$$Y_{ti} = \mu_t + U_{0i} + R_{ti} \quad , \quad (3.6)$$

where Y_{ti} is the outcome response variable of individual i at time t , $U_{0i} \sim N(0, \sigma_o^2)$ which explains the variability of the fixed part μ_t and $R_{ti} \sim N(0, \sigma^2)$ is the random part.

To fit this model, we note that, the fixed part does not contain a constant term, but is based on m dummy variables for the m measurement occasions. We let d_{hti} be m dummy variables defined for $h=1,2,...,m$ by

$$d_{hti} = \begin{cases} 1, & t=h \\ 0, & t \neq h \end{cases}$$

Therefore, the fixed part μ_t in (3.6) above can be written as

$$\mu_t = \sum_{h=1}^m \mu_h d_{hti} \quad (3.7)$$

and the model is formulated as

$$Y_{ti} = \sum_{h=i}^m \mu_h d_{hti} + U_{0i} + R_{ti} \quad (3.8)$$

where the symbols are as above.

Extending the Maxwell and Delaney (1990) model represented by equation (3.8), an appropriate model for analysing data with more parameters can be got.

Incase of our worms control data to be analyzed in the next chapter, there are relevant explanatory variables included in the fixed part. i.e. there are individual independent explanatory variables z_k , $k=1,2,\dots,q$. These variables may represent; sex, educational level etc. Where these variables are categorized, they can be represented by dummy variables $[0,1]$.

These individual independent variables are level 2 variable in the multilevel approach, and they are called between subjects variables in repeated measurements analysis. This is because each subject is either in one group of these variable or the other. For instance, if we consider sex, each individual is either in one group or the other. i.e. the pupil is either a boy or a girl.

Again we have numerical variable describing the measurement occasion denoted by $s(t)$ which could be the elapsing time between measurement occasions. Also there is interaction of between subject and within subject variables. i.e. the interaction effect of the main effects-the within between interactions which are cross-level interactions. Within subject variables here, describes the observations on each trials or measurements occasions. Therefore, the more appropriate model where the data have all above variables is given by

UNIVERSITY OF NAIROBI
CHIROMO LIBRARY

$$Y_{ti} = \sum_{k=1}^q \alpha_k z_{ki} + \sum_{k=1}^q \lambda_k z_{ki} s(t) + \sum_{h=1}^m \mu_h d_{hti} + U_{0i} + R_{ti} \quad (3.9)$$

where

μ_h is the observation on h^{th} measurement occasion.

z_k is the background information on i^{th} individual.

α_k is the effect due to z_{ki} (main effect).

$s(t)$ is the elapsing time to the next measurement occasion

λ_k is the interaction effect due to z_{ki} and $s(t)$

Its assumed here, that the level 1 residual R_{ti} have population means 0 and variance σ^2 , given the values of explanatory variables. Therefore, $R_{ti} \sim N(0, \sigma^2)$ where level 1 residual arise due to variability of within subject variables. Also its assumed that the vector $(U_{0i}, \dots, U_{qi})$ which describes the variability of between subject variables is independent of the level 1 residual R_{ti} and that all residuals have population means 0.

The vector $(U_{0i}, \dots, U_{qi})$ has a multivariate normal distribution with a constant covariance matrix where,

$$\text{var}(U_{hi}) = \sigma_{hh} = \sigma_h^2 \quad (h=1, 2, \dots, q)$$

$$\text{cov}(U_{hi}, U_{ki}) = \sigma_{hk} \quad (h \neq k=1, 2, \dots, q)$$

The model therefore, implies that for complete data vector, all variances are equal and also all covariances are equal.

The expression for the covariance matrix of the complete data vector associated with model represented by equation (3.6), conditional on the explanatory variables is the $m \times m$ matrix

$$V(Y) = \begin{pmatrix} \sigma_0^2 + \sigma^2 & \sigma_0^2 & \cdot & \cdot & \cdot & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 + \sigma^2 & \cdot & \cdot & \cdot & \sigma_0^2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_0^2 & \sigma_0^2 & \cdot & \cdot & \cdot & \sigma_0^2 + \sigma^2 \end{pmatrix}$$

In this covariance matrix, all residual variances are the same and all residuals within subject correlations are equal.

$$\rho_i = \rho(Y_{ti}Y_{si}) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} \quad (3.10)$$

It should be noted that, this model is very restrictive and therefore unlikely. e.g. Where measurements are ordered in time, correlation is often larger between nearby measurements than between measurements that are far apart resulting in complex covariance structure.

3.5 The logistic model

Anderson(1972) has widely used logistic regression to model association between binary response and explanatory variables. Binary data arise in situation where the response can take only one of the two possible values conveniently denoted by 0 and 1. These observations are common in medical trials.

For instance, after observing children, we may conclude, the presence of worms (1), or the absence of worms (0) in each child. Where the response assumes the probabilities of failure and successes respectively (Bernoulli distribution).

In most investigations, a vector of covariates or explanatory variables $(x_1, x_2, \dots, x_p)$ is associated with each individual. This vector consists of measured variable thought likely to influence the probability of a positive response. The objective of analysis therefore is to investigate the relationship between the response probability $\pi = \pi(x)$ and the explanatory variables, $x' = (x_1, x_2, \dots, x_p)$, often of course, a subset of the x_s is of primary importance.

In generalized linear models, the issue is to describe the proportion of successes in each subgroups in terms of factor levels and other explanatory variables which characterize the subgroup.

This is done by modeling the probabilities π_i as $g(\pi_i) = x'_i \underline{\beta}$

where x_i is a vector of explanatory variables and $\underline{\beta}$ is a vector of parameters and g the link function.

since the simplest model $\pi = x'_i \underline{\beta}$ has disadvantages that although π is the probability, the fitted values $x'_i \underline{b}$ may be outside the interval $[0,1]$.

we model π using a cumulative probability distribution.

When the response is rare, efficiency is enhanced by taking separate samples from the subpopulations of responders and non-responders. Thus case control studies in epidemiology typically samples most of the disease cases that occur during a specified calender period.

Suppose Y_{it} is binary response variable; defined as 1 if the individual i has a particular worm at measurement occasion t and 0 otherwise. Then Y_{it} has a Bernoulli probability distribution with

$$P_r(Y_{it} = y) = p_{it}^y (1 - p_{it})^{1-y} \quad (3.11)$$

Where p_{it} is the probability of having a specific worm by individual i in measurement occasion t . The relation between p_{it} and the explanatory variables is made through a suitable linear predictor and the logit link function. This transforms the linear predictor of explanatory variables, which may have a value between plus and minus infinity to a probability which necessarily lies between zero and one.

Using the logit link function $\log[p_{it}/(1 - p_{it})]$, the simple logistic regression model is

$$\log[p_{it}/(1 - p_{it})] = \underline{\beta}' \underline{x}_{it} \quad (3.12)$$

where $\underline{\beta}' \underline{x}_{it} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$. Equation (3.12) can be rewritten as

$$p_{it} = \exp(\underline{\beta}' \underline{x}_{it}) / [1 + \exp(\underline{\beta}' \underline{x}_{it})] \quad (3.13)$$

and the model including residual heterogeneity as

$$p_{it} = \exp(\beta' \underline{x}_{it} + e_{it}) / [1 + \exp(\beta' \underline{x}_{it} + e_{it})] \quad (3.14)$$

where $\underline{x}_{it}$ is a vector of explanatory variables, β is a vector of unknown parameters and e_i is an individual specific term summarizing the effect of the omitted variable.

Residual heterogeneity result due to omitted variables in some individuals.

3.6 Estimation of parameters for logistic model

The maximum likelihood estimates of the $p \times 1$ parameters vector β is obtained by maximising the log likelihood equation (3.12).

Let Y_{ij} , $j=1,2,\dots,n_i$; $i=1,2,\dots,k$, represent the j^{th} measurement on the i^{th} object. there are n_i measurement on subject i and $\sum_{i=1}^k n_i$ total measurements.

Correlated data are modeled using the link function and linear predictor set up.

Let the vector of measurements on the i^{th} subject be $\underline{Y}_i = [Y_{i1}, Y_{i2}, \dots, Y_{in}]'$

with corresponding vector of means $\underline{\mu}_i = [\mu_{i1}, \mu_{i2}, \dots, \mu_{in}]'$ and let v_i be an estimate of the covariance matrix of $\underline{Y}_i$. The generalized estimating equation is given as

$$\sum_{i=1}^k \frac{\delta \mu'_i}{\delta \beta} v_i^{-1} [\underline{Y}_i - \underline{\mu}_i(\beta)] = 0 \quad (3.15)$$

For binary data , Diggle, Liang and Zeger(1994) pointed out that, the modeling association among binary response with correlation has a disadvantage, and they proposed using odds ratio instead.

Therefore for binary data, the correlation between the j^{th} and k^{th} response is by definition

$$corr(Y_{ij}Y_{ik}) = \frac{P_r(Y_{ij} = 1, Y_{ik} = 1) - \mu_{ij}\mu_{ik}}{\sqrt{[\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})]}} \quad (3.16)$$

The correlation, therefore, is constrained to be within limits that depend on a complicated way on the means of the data.

The odds ratio, defined as

$$OR(Y_{ij}, Y_{ik}) = \frac{P_r(Y_{ij} = 1, Y_{ik} = 1) \times P_r(Y_{ij} = 0, Y_{ik} = 0)}{P_r(Y_{ij} = 1, Y_{ik} = 0) \times P_r(Y_{ij} = 0, Y_{ik} = 1)} \quad (3.17)$$

is not constrained by the means and is preferred for our data, which is binary in nature.

Carey ,Zeger and Diggle (1993) proposed a method for fitting log odds ratio as

$$Log(OR(Y_{ij}, Y_{ik})) = \underline{z}_{ij}'\underline{\alpha} \quad (3.18)$$

where $\underline{z}_{ijk}'$ is a vector of covariates and $\underline{\alpha}$ is a vector of association parameters to be estimated.

There are several advantages for adopting the above GEE modeling approach for our data. first, the GEE estimating algorithm is computationally efficient for relatively large data sets such as ours (where $n=117$ and $m=3$). It is also widely implemented in standard-use statistical packages for correlated data analysis such as SAS, SPSS and STATA.

Third, we note that the model depicted in equation (3.17) uses odds ratio rather than correlation coefficients to model dependence between observations on different occasions from the same child. This is preferable since odds ratios provide a more natural framework for modeling the within-subject dependencies that are induced by the binary response we have in this study.

3.7 Probit models

One of the original models used for bioassay data is called the probit model.

This model gives numerical results very much like the logistic model described above.

In this model, the link function is the inverse cumulative normal distribution. i.e. ϕ^{-1} where

$$\phi = \frac{1}{\sqrt{(2\pi\sigma^2)}} \int_{-\infty}^x \exp \frac{-1}{2} \left\{ \frac{x - \mu}{\sigma} \right\}^2 ds \quad (3.19)$$

which is equivalent to

$$\phi \left\{ \frac{x - \mu}{\sigma} \right\} \quad (3.20)$$

where ϕ denote the cumulative probability function for the standard normal distribution i.e. $N(0,1)$ we therefore have

$$\phi^{-1}(\pi) = \beta_1 + \beta_2 x = \frac{x - \mu}{\sigma} \quad (3.21)$$

That is $\beta_1 = \frac{-\mu}{\sigma}$ and $\beta_2 = \frac{1}{\sigma}$

Probit models were introduced by Chester I. B., et al in 1935, and assume that

$$Pr(Y = 1/X = x) = \phi(x'\beta) \quad (3.22)$$

where X is vector of regressions and Y is binary outcome variate generated by a simple latent variable model. Suppose that

$$Y^* = x'\beta + \epsilon \quad (3.23)$$

where $\epsilon/x \sim N(0, 1)$ and Y^* is an indicator for whether the latent variable Y^* is positive.

$$Y = 1(Y^* > 0) = \begin{cases} 1, & \text{if } Y^* > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3.24)$$

then it is easy to show that

$$Pr(Y = 1/X = x) = \phi(x'\beta) \quad (3.25)$$

It should be noted that, the probit function is the inverse cumulative distribution associated with the standard normal distribution which has applications in exploratory graphs and specialized regression modeling of binary response variables. For standard normal distribution $N(0,1)$ the cumulative distribution function is the one denoted by $\phi(x)$ which is a continuous increasing function whose domain is the real line and range is $[0,1]$.

The multivariate probit model is a generalization of the probit model which is used to estimate several correlated binary outcomes jointly. e.g. presence or absent of two types of worms. where both variables are binary, then the multivariate probit model would be appropriate.

e.g. if Y_1 and Y_2 are 2 binary dependent variables.

$$\begin{cases} y_1 = 1(Y_1^* > 0) \\ y_2 = 1(Y_2^* > 0) \end{cases} \quad (3.26)$$

with

$$Y_1^* = x\beta_1 + \epsilon_1 \quad \text{and} \quad Y_2^* = x\beta_2 + \epsilon_2 \quad (3.27)$$

and

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} | x \sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\} \quad (3.28)$$

Hence the bivariate probit model estimates the values of β_1 , β_2 and ρ .

3.8 Maximum likelihood estimation for the probit model

This method involves specifying the joint probability distribution function for the sample data.

Assuming that we have a random sample of size n drawn from some probability distribution, we let $Y_1, Y_2, \dots, Y_n$ denote the random variable that we sample. If Y_i are discrete as the case for our data, then the probability that Y_i takes any particular value y_i is given by the probability mass function

$$\text{Prob}(Y_i = y_i) = f(y_i) \quad (3.29)$$

However $f(y_i)$ will depend upon some parameters $\beta_1, \dots, \beta_k$ which we want to estimate. To indicate the dependence of the probabilities on these parameters, we write

$$f(y_i) = f(y_i/\beta_1, \dots, \beta_k) \quad (3.30)$$

The joint probability mass for the random variables $Y_1, \dots, Y_n$ is then the product of the individual probability mass functions.

$$Prob(Y_i = y_i, \dots, Y_n = y_n) = \prod_{i=1}^n f(y_i/\beta_1, \dots, \beta_k) \quad (3.31)$$

This is a function of both the values $y_1, \dots, y_n$ taken by the random variables and the parameters $\beta_1, \dots, \beta_k$. However, once a sample has been drawn, the values taken by the random variables are known. Therefore for any choice of the parameters $\beta_1, \dots, \beta_k$, we can evaluate the probability of obtaining the observed sample values.

$$L(\beta_1, \dots, \beta_k) = \prod_{i=1}^n f(y_i/\beta_1, \dots, \beta_k) \quad (3.32)$$

where $Y_1, Y_2, \dots, Y_n$ are now fixed at their observed values.

For our probit model we let Y_i denote the dependent variables and $x_{1i}, \dots, x_{ki}$ denote the independent variables. A reasonable choice of functional form for the probabilities is

$$Prob(Y_i = 1/x_{1i}, \dots, x_{ki}) = F(\beta'x_i) \quad (3.33)$$

where $\beta'x_i = \beta_1 x_{1i} + \dots + \beta_k x_{ki}$ and $F(\cdot)$ is a non decreasing function such that

$$\lim_{t \rightarrow -\infty} F(t) = 0$$

$$\lim_{t \rightarrow +\infty} F(t) = 1$$

The cumulative normal distribution will satisfy these condition and that is why it is used in probit model i.e.

$$F(t) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^t e^{-\frac{u^2}{2}} du = \phi(t) \quad (3.34)$$

In this case, if $\beta'x_i$ is large, then the probability that Y_i equals to one is close to one.

Similarly, if $\beta'x_i$ is small, then the probability that Y_i equals one is close to zero. Whatever values the independent variables takes, the probability that Y_i equals one will be admissible. i.e. between zero and one

The log likelihood in this case is given by

$$\text{Log}L(\beta_1, \dots, \beta_k) = \sum_{i=1}^n y_i \text{Log}(\phi(\beta'x_i)) + \sum_{i=1}^n (1 - y_i) \text{Log}((1 - \phi(\beta'x_i))) \quad (3.35)$$

It should be noted that the probit log likelihood are globally concave and hence relatively easy to maximize using the newton raphson algorithm.

For our binary data, the procedure of estimation and prediction involve first computing maximum likelihood probit estimates, next creating dummy variables which equals one if the case is predicted correctly and equals zero otherwise. Finally, we calculate the percentages of successes. The number produced is the "percentage correctly predicted".

Chapter 4

DATA ANALYSIS

The data for 117 pupils was analyzed by fitting both the logit and probit models. Prior to fitting these models, summary statistics were carried out. These are given in Table 4.1 which shows the number of children from the sample population who were infected by different types of worms at different visit. We observe that, after the first visit, the number of children

Table 4.1: Worms observed

Visit	Ascaris	hookworms	Trichuris	Schistosomula
Baseline	30(25.64%)	64(55.70%)	63(53.85%)	105(89.74%)
Four months	14(12.61%)	31(27.93%)	40(36.04%)	65(58.56%)
Eight months	19(17.59%)	34(31.19%)	48(44.04%)	94(86.24%)

infected by the four types of worms, dropped by the first four months i.e. second visit and slightly increased by the next four months i.e. the third visit.

Tables 4.2-4.5 give the results from the logit and probit analyses. These describe the effect of demographic factors that are associated with different worm infection.

Table 4.2: Ascaris model

Characteristic	LOGIT LINK			PROBIT		
	estimate	s.e.	95% c.i	estimate	s.e.	95% c.i
Intercept	-1.077	0.500	-2.073-0.080	0.640	0.278	-1.186-0.095
gender						
-boys	-ref					
-girls	0.535	0.405	0.259-1.329	0.312	0.222	-0.123-0.747
Educational level						
-class 5	-ref					
-class 6	0.301	0.421	-0.524-1.125	0.163	0.227	-0.281-0.608
Supplement						
-no	-ref					
-yes	-0.927	0.395	-1.702-0.153	-0.511	0.218	-0.938-0.085
Treatment						
-no	-ref					
-yes	0.819	0.427	-0.017-1.656	0.437	0.233	-0.019-0.893
hookworms						
-no	-ref					
-yes	-0.893	0.397	-1.670-0.116	0.494	0.216	-0.917-0.071
Trichuris						
-no	-ref					
-yes	-0.960	0.383	-1.711-0.209	-0.521	0.210	-0.932-0.110
Schistosomula						
-no	-ref					
-yes	-0.242	0.514	-1.249-0.765	-0.152	0.270	-0.680-0.377

Table 4.2 shows that infection by ascaris was significantly associated with gender, presence of hookworms and trichuris. The infection on girls was higher by 70% ($= (1 - \exp(0.535)) \times 100\%$) than in boys. On education level, pupils in class six infection is higher by 40% ($= (1 - \exp(0.301)) \times 100\%$) than those in class five. However those given the macronutrient supplement are less likely to be infected by 40% and also those treated for previous infection, were 2.3 times higher likely to be infected, though non-significantly.

Table 4.3: Hookworms model

characteristic	LOGIT LINK			PROBIT		
	estimate	s.e.	95% c.i	estimate	s.e.	95% c.i
intercept	0.862	0.504	-0.126-1.849	0.517	0.304	-0.079-1.112
gender						
-boys	-ref					
-girls	-0.540	0.351	-1.228-0.240	-0.328	0.209	-0.738-0.081
Educational level						
-class 5	-ref					
-class 6	-0.408	0.331	-1.057-0.240	-0.245	0.197	-0.632-0.142
Supplement						
-no	-ref					
-yes	-0.363	0.340	-1.029-0.304	-0.207	0.203	-0.604-0.190
Treatment						
-no	-ref					
-yes	0.821	0.331	0.172-1.471	0.498	0.201	0.104-0.893
Ascaris						
-no	-ref					
-yes	-0.878	0.395	-1.651-0.104	-0.537	0.240	-1.008-0.066
Trichuris						
-no	-ref					
-yes	-0.442	0.317	-1.063-0.180	-0.269	0.191	-0.643-0.104
Schistosomula						
-no	-ref					
-yes	-0.854	0.398	-1.634-0.074	-0.514	0.228	-0.961-0.068

Table 4.3 shows that hookworms infection was significantly associated with gender, treatment, presence of ascaris and schistosomula. Unlike the ascaris analysis, girls are less likely to be infected by hookworms by 40% than boys. Those in class six are less likely to be infected by 30% than those in five and supplement reduces infection by 30% though not significantly. Those treated for previous infection were 2.3 times likely to be infected.

Table 4.4: Trichuris model

characteristic	LOGIT LINK			PROBIT		
	estimate	s.e.	95% c.i	estimate	s.e.	95% c.i
intercept	0.551	0.526	-0.479-1.582	0.340	0.323	-0.294-0.973
gender						
-boys	-ref					
-girls	0.426	0.359	-1.457-0.950	0.153	0.221	-0.280-0.585
Educational level						
-class 5	-ref					
-class 6	-0.080	0.367	-0.800-0.639	-0.408	0.226	-0.491-0.395
Supplement						
-no	-ref					
-yes	0.579	0.354	-0.116-1.273	0.357	0.217	-0.069-0.784
Treatment						
-no	-ref					
-yes	0.088	0.349	-0.597-0.773	0.053	0.216	-0.369-0.476
Ascaris						
-no	-ref					
-yes	-0.940	0.390	-1.705-0.175	-0.583	0.239	-1.050-0.115
hookworms						
-no	-ref					
-yes	-0.444	0.313	-1.058-0.170	-0.273	0.193	-0.651-0.106
Schistosomula						
-no	-ref					
-yes	-0.429	0.320	-1.506-0.199	-0.258	0.195	-0.639-0.124

Table 4.4 shows that trichuris was significantly associated with supplement and presence of ascaris. Girls are more likely to be infected by 50% than boys while class six pupils are 80% less likely to be infected than pupils in class five. Those given supplement were more likely to be infected by 58% than those not given. While those who were treated for previous infections were 5% more likely to be infected.

Table 4.5: Schistosomula model

characteristic	LOGIT LINK			PROBIT		
	estimate	s.e.	95% c.i	estimate	s.e.	95% c.i
intercept	2.179	0.633	-0.938-3.420	1.254	0.355	0.558-1.951
gender						
-boys	-ref					
-girls	-0.867	0.316	-1.486-0.249	-0.517	0.178	-0.866-0.168
Educational level						
-class 5	-ref					
-class 6	-0.110	0.334	-0.766-0.545	-0.050	0.197	-0.436-0.337
Supplement						
-no	-ref					
-yes	0.161	0.341	-0.507-0.829	0.079	0.194	-0.302-0.459
Treatment						
-no	-ref					
-yes	1.920	0.485	0.969-2.870	1.080	0.236	0.618-1.543
Ascaris						
-no	-ref					
-yes	-0.168	0.502	-1.152-0.816	-0.101	0.280	-0.651-0.448
hookworms						
-no	-ref					
-yes	-0.886	0.409	-1.687-0.085	-0.450	0.225	-0.890-0.010
Trichuris						
-no	-ref					
-yes	-0.441	0.332	-1.092-0.209	-0.282	0.193	-0.659-0.095

Table 4.5 shows that schistosomula infections was significantly associated with gender, treatment and presence of hookworm, girls are less prone by 60% and those in class six were slightly less prone by 10% and those given supplement are more infected by 20% though not significantly. Those treated for previous infections were highly likely to be infected 6.8 times more.

Chapter 5

CONCLUSION

Among the class five and six school going pupils from Bondo district studied, it was observed that the infection dropped with time for all the four types of worms namely; ascaris, hookworms, trichuris and schistosomula and there after increased slightly.

In general, it can be concluded that educational level have little effect and therefore insignificant. This could be due to the fact that the class levels choosen are too close and therefore children age and other control factors could be almost the same. Also sex have little effect and therefore insignificant in all the models. It implies that both girls and boys have almost equal chances of being infected with worms.

Supplement and treatment can be applied in control. However other measures needs to be devised for full control of the above four types of worms.

The study reveals that the infections are not related and therefore it can be recommended that each type of worm need to be controlled independently.

In all the analysis carried out, the probit model gave a better fit as it produced smaller standard errors for all parameter estimates than the logit model.

5.1 Recommendation for further study

In the analysis carried out on these data, association between worms was determined by including worm types as covariate in the model. For instance, in modelling ascaris, hookworm, trichuris and schistosomula were included as covariates. We know that indeed worm types may be dependent. It may therefore be more appropriate to treat two worm types as dependent variables in a joint model. These models may therefore be more appropriate and need to be investigated as a further analysis tool for this data set.

5.2 REFERENCES

1. Barnhart H.X., and Williamson J.m., 1998. Goodness-of-fit tests for GEE modeling with binary data. *Biometrics*. 54. p720-729.
2. Bliss C.I., 1934. The method of probit science. 79. p38-39.
3. Carry V., Zeger S. L., and Diggle P., 1993. Modeling multivariate binary data with alternating logistic regression" *Biometrika*. p517-526.
4. Collett. D., 1991. Modeling Binary Data Chapman and Hall/CRC. p56.
5. Diggle P. J. , Liang K.Y. and Zeger S.L.,1994. Analysis of Longitudinal Data. Oxford: Oxford science. p52-55.
6. Everitt S.K. Statistical methods for medical investigation O.U.P. New York: U.S.A.
7. Finney D.J. ,1947. Probit analysis (1st edition) Cambridge University Press: Cambridge. U.k.
8. Finney D.J., 1947. Probit analysis (3rd edition) Cambridge University Press: Cambridge. U.k.
9. Hedeker D., and Gibbons R.D. Longitudinal data analysis.
10. Liang K.Y. and Zeger S.L.,1986 "Longitudinal data analysis using generalized linear models" *Biometrika*. p657-671.
11. Nokes C. , Grantham M. S., Sanyer A.W., Cooper E.S. and Bundy D.A.P., 1992. "Parasitic helminth infection and cognitive function in school children"
12. Thall P.F. and Vail S.C. , 1990. "Some covariances models for longitudinal count data with over dispersion" *Biometrika*. p657-671
13. Tom A.B.S. and Roel J.B. , 1999 Multilevel analysis (An introduction to basic and advanced multilevel Modeling)
14. Zeger S. L.and Liang K.Y. ,1986 "Longitudinal data analysis for discrete and continuous outcomes" *Biometrics*. p121-130.