Nonlinear Mixed Effects Modeling of Gametocyte Carriage in Patients with Uncomplicated Malaria

G B Distiller

Department of Statistical Sciences
University of Cape Town

Supervisors: Dr Francesca Little & Dr Karen Barnes

May 16, 2007

Acknowledgements

I would like to thank Dr Francesca Little for her excellent supervision and guidance and Dr Karen Barnes for her expert knowledge and constructive feedback. I would also like to thank Professor Theo Stewart for helping me getting to grips with using LaTeX.

I would also like to acknowledge my wife Suki Goodman for her continued support and understanding.

Abstract

Malaria is the most important parasitic disease today with hundreds of millions of people being infected each year, mostly in sub-saharan africa. Malaria is caused by *Plasmodium* parasites which are transmitted by a mosquito vector. Human infectivity is the probability of a mosquito becoming infected after biting an infected person, and plays a crucial role in the spread of an infection.

The malaria parasites cause disease in the asexual form and a proportion of these asexual parasites eventually evolve into a sexual form known as gametocytes. A mosquito then gets infected if she feeds on a person with mature male and female gametocytes in their bloodstream. While there are several factors that influence host infectivity, the density of gametocytes appears to be the best single measure that is related to the human host infectivity to mosquitoes. A log-sigmoid curve has been estimated to represent the relationship between gametocyte density and infectivity.

The mathematical modeling of malaria began approximately a century ago with Ronald Ross's model that was based on a couple of differential equations that showed the changes in the number of infected hosts and vectors. This model produced some profound insights despite its simplicity and the many assumptions that it made. Since then, this model has been extended by many different researchers who have attempted to incorporate an array of extra complexities.

Despite the obviously important role that gametocytes play, none of the models reviewed had attempted to directly model gametocyte distributions over time. Gametocyte carriage has typically been estimated from the more frequently measured asexual parasite density, usually using a simple mechanism such as a proportional gametocyte switching rate. The objective of this dissertation is to model observed gametocyte densities over time for relevant patient profiles. The contribution that this research aims to make is therefore focused on the human infectivity component of models for the transmission dynamics of malaria.

Nonlinear mixed effects modeling was the method selected as most appropriate in order to meet this objective. This is due to the obviously nonlinear shape of a typical gametocyte density-time profile, as well as the fact that there is a large amount of individual variation in these profiles. Lastly, mixed effects modeling can handle a situation where there is sparse data on some patients by 'borrowing' information from the larger group.

The modeling strategy began by attempting to find a nonlinear function that could track the underlying data-generating mechanism. A variety of different functions were examined and the one that was finally selected was a modified

version of the critical exponential function. This was chosen due to it's robustness across the different datasets as well as the fact that it appeared to be flexible and capable of assuming a variety of different shapes. This function behaves similarly to an ordinary exponential decay function but with the first term being dependent on time and hence not staying constant.

The data used are from a study that used measurement points that were based around the expected clinical and asexual parasitological response to malaria treatment. A design that is more optimal for the measurement of gametocytes would improve the accuracy and richness of the data, and allow more complex functions to be examined. While the Double Fourier is one such function, there are questions around the importance of capturing a wave-like pattern in the data seeing as the process of sequestration can lead to the observed gametocyte densities not being reflective of the true gametocyte burden. Furthermore the wave-like structure would not be observed for all patients unless they all carried synchronous infections. It could therefore make more sense to estimate curves that 'smooth' through these waves, much like those observed with this data.

The second stage of modeling involved examining a number of patient and disease-specific covariates in an attempt to explain some of the variation between individuals. These included a variable that classified patients into one of three categories based on the observed shape of their gametocyte density-time profile. Age, site, initial asexual parasite density (logged to the base 10), and the observed patient category were the covariates that were found to improve the model.

Various curves for different patient profiles (according to the covariates that were found to be important for the model) were then generated from the model. These curves provide information that could be used at each iteration in a larger model of the entire malaria transmission cycle to firstly predict gametocyte density in the population, and secondly to estimate the average level of infectivity at a community level from this density. The infectivity estimate would be used as an input for the next iteration.

It was surprising that the antimalarial drug resistance mutation variable did not enter the model and it is conjectured that this is related to power and the fact that there were very few patients who had resistant infections and failed treatment. In addition to this, the fact that most treatment failures are rescued and removed from the study makes it very difficult to accurately estimate their subsequent gametocyte density-time profiles. Either a questionable assumption (such as equating their curves with treatment successes) has to be made, or better data that tracks their gametocytes after being withdrawn from the study need to be obtained.

Lastly, it is pertinent that this research is applicable only to malaria patients who carry gametocytes. A full model of the dynamics of malaria transmission

and the spread of resistance would need to incorporate an additional component that captures the determinants of gametocyte prevalence i.e. what causes some people to exhibit gametocytes while others do not.

Contents

List of Figures

List of Tables

Chapter 1

Literature Review

1.1 Introduction

This literature review begins with an introduction to the modeling of infectious diseases in general, and malaria in particular. The characteristics that are specific to malaria are then discussed with specific emphasis on how they relate to the carriage of gametocytes. The third section reviews previous malaria models and ends by presenting a possible alternative model that is driven by gametocytes.

1.2 Modeling Infectious Diseases

Despite the fact that many of the infectious diseases that have afflicted humans in the past are gradually disappearing from the modern world, there are still many diseases that affect hundreds of millions of people (Bailey, 1975; Becker, 1989). While medicine today is usually effective at curing many infectious diseases once they have been contracted, the best results have been experienced with prevention (Bailey, 1975).

It is obviously not ethical to deliberately infect people in order to assess an intervention (Becker, 1989). Appropriate quantitative modeling can therefore be of immediate benefit to epidemiology and public health by allowing hypotheses to be tested, biological parameters to be estimated, and different intervention strategies to be assessed (Bailey, 1975; Smith et al., 2006). Mathematical models provide a mechanism of moving from the description of the role that an infected individual plays, to a description of how the disease spreads through a population over time (Becker, 1989; Pongtavornpinyo, 2006).

Thompson (2000) calls a model "...a mathematical summary of our best guess as to what is going on in a part of the real world". Becker (1989) says that the objective of modeling epidemic data is to improve our understanding

of infectious diseases and how they spread. Molineaux (1985) speaks of using epidemiological models in training, for planning control, and for research.

When modeling infectious diseases, there is typically interest in the transmission and spread of the particular disease as well as trying to understand and predict the outbreak of epidemics. Where a disease is endemic, attention is usually paid to how the level of endemicity is related to factors that can be controlled through public-health policies and interventions. (Bailey, 1975)

The more detail that a model has included, the more assumptions about interactions it will have made, and this will greatly increase the chance of making a wrong critical assumption (Koella, 1991). It is therefore important to strive for a model that is simple yet realistic, and accurate enough to be of use in decision making. A simple model with an adequate fit to the data can often reveal what the important characteristics are with greater clarity then a more complex model (Becker, 1989; Smith et al., 2006). Koella (1991) suggests that the increasing complexity of the models published during the 1960's and 70's could explain why mathematical models struggled to gain general acceptance despite the early recognition by people like Ross and MacDonald of the importance of quantifying malaria dynamics.

The aim of malaria epidemiology is to understand the dynamics of the disease in order to be able to build and manage more efficient control programs (Koella, 1991). Biological field studies have helped build knowledge about malaria but cannot always explain the discrepancies observed between malaria patterns in different places. Mathematical epidemiology can be used to explore these differences by integrating many factors into a single coherent picture and thus relate the various factors of the transmission cycle to each other and to relevant biological characteristics of the mosquito (MacDonald, 1957; Koella, 1991). This complex interaction of various factors leads to nonlinear terms in the description of malaria transmission and it is virtually impossible to grasp the effect of such nonlinearities without a mathematical model (Koella, 1991).

Ronald Ross is widely credited with being the first person to attempt to develop a structured mathematical theory for the study of epidemiology (Heesterbeek, 2002; Fine, 1975). This introduction ends with the following famous quote by Ross (1911): '...As a matter of fact all epidemiology, concerned as it is with the variation of disease from time to time or place to place, must be considered mathematically, however many variables are implicated, if it is to be considered scientifically at all. To say that a disease depends on certain factors is not to say much, until we can also form an estimate as to how largely each factor influences the whole result...'.

1.2.1 History

In the 17th century, John Graunt and William Petty studied the London Bills of Mortality and this work may be taken to mark the beginning of medical statistics. Unfortunately there was no coherent epidemiological theory of what drives epidemics and it took nearly 200 years before any real progress in the biological field was achieved. Consequently, mathematical theories began to be developed only after a clear biological basis for the cause of infectious diseases was established. However since the turn of the 20th century there has been continued acceleration in this area of research and according to Bailey (1975) "... the advancement already achieved by mathematical investigation in biological subjects such as evolution and genetics is extremely encouraging". (Bailey, 1975)

Most of the work on infectious diseases in the early 1900's was deterministic in character i.e. did not incorporate probability aspects into the processes studied. In 1906 Hamer looked at the dependence of an epidemic on the number of susceptible individuals and the rate of contact between these susceptibles and the infectious individuals. The simple mathematical assumptions that he made are basic to all subsequent deterministic models. (Bailey, 1975)

Ronald Ross discovered that malaria is spread by mosquitoes in 1898 and got the Nobel prize in 1902 for this discovery (Heesterbeek, 2002; Fine, 1975). In 1908 Ross produced a report wherein he attempted to specify and tie together all the major factors involved in the transmission of malaria. He did this using a simple algebraic equation. He began to generalise this approach in 1911 and derived a system of difference equations that estimated the incidence and prevalence patterns for various scenarios. The final development in his mathematical theory happened in 1916 and took the form of differential equations. (Fine, 1975) This led to the first well-organised mathematical theory that could be used as a research tool in epidemiology (Heesterbeek, 2002; Bailey, 1975).

Prior to this, most research involved fitting curves to epidemic data and then comparing the estimated results with observed data. This approach did not begin with making assumptions about the mechanism of transmission. This latter approach was dubbed *a priori* by Ross and has become the standard approach used by many epidemiological modellers (Fine, 1975).

From 1927-1939 more advanced research of the same kind was conducted by Kermack and McKendrik culminating in the celebrated Threshold Theorem (see section 1.2.3 for more details). Soper carried out further deterministic work involving measles in 1929. (Bailey, 1975)

Chance and random variation became more important as the extent of available epidemiological data grew, and the need to occasionally study much smaller groups like households arose. In 1926 McKendrick published the first stochastic

treatment of an epidemic process by taking the *probability* of a new case as being proportional to the numbers of both susceptibles and infectious cases as opposed to the actual number of new cases used in the deterministic models. This is a 'continuous-infection' model whereby an individual is infectious from the moment the infection is received until recovery or death. Unfortunately this paper did not attract much attention and similar research was only picked up again twenty years later. (Bailey, 1975)

In 1931 Greenwood treated the number of cases occurring at any stage as having a binomial distribution and this led to a model with a chain of binomial distributions termed a 'chain-binomial' model. Reed and Frost were also involved in similar work that they were using in lectures and discussions in the U.S.A. (Bailey, 1975)

The 40's and 50's saw further work being done in both deterministic and stochastic areas, and included Whittle's derivation of a stochastic version of the threshold theorem (Heesterbeek, 2002; Bailey, 1975). Stochastic models managed to model the outbreak of waves of measles more realistically than their deterministic predecessors and there was a renewed focus on both continuous infection and chain-binomial models. (Bailey, 1975)

After 1957 the amount of research in this area exploded, and more recently advances in computer technology have enabled models with greater complexity to be developed (Bailey, 1975). Dietz (1988) speaks about the fact that despite over 75 years of development in malaria mathematical epidemiology, the subject is very much at its beginning.

1.2.2 Epidemiological principles

This section presents the general biological process that underlies the spread of most infectious diseases using examples that are specific to malaria.

An infectious disease can be defined as one that is capable of being transmitted from an infected host to an uninfected susceptible. The picture of transmission gets more complicated when an intermediate host or ***vector*** is involved as in the case of malaria where the parasite transmits from man to mosquito and from mosquito back to man again. Consequently there are two different populations of infectives and susceptibles involved i.e. human hosts and mosquito vectors. (Bailey, 1975)

An individual is exposed to an infection of some sort i.e. a susceptible person gets bitten by a mosquito infected with the malaria parasite. This individual could have a degree of innate protection depending on his/her physiological defences that may or may not be acquired due to previous exposure. (Bailey, 1975) This is called ***immunity*** and with malaria is usually linked to the intensity of

transmission i.e. people in endemic areas gain immunity with increasing age.

Alternatively, the infection may manage to establish itself and complete its life cycle. Usually this involves a ***latent period*** where the development does not emit any infectious material. After the latent period ends, infections can be communicated to a susceptible by the infected individual who is termed ***infective***. (Bailey, 1975) In the case of malaria, the latent period encompasses the early part of the parasite's life cycle and infectious material is only emitted (to a susceptible mosquito) once the parasite has evolved into its sexual form called gametocytes. The ***infectious period*** is the length of time that a person remains ***infective*** for (Bailey, 1975) i.e. the length of time that a person carries gametocytes and can hence infect susceptible mosquitoes. There is considerable variation in the duration of this time period (Barnes and White, 2005).

Symptoms may occur at some stage of the infection. If this happens during the latent period then isolating the individual will stop the infection spreading. Unfortunately it is more common for symptoms to occur after the start of the infectious period thereby making it more difficult to prevent the spread of the infection. The window of time for transmission will be the time from the start of the infectious period to the time of isolation. (Bailey, 1975) With malaria the symptoms present during the latent period but it is difficult to effectively isolate an infected person as they need to be isolated from any mosquitoes (as opposed to susceptible humans) for the entire duration that they carry gametocytes in their bloodstream.

The period between receiving an infection and the appearance of symptoms is called the ***incubation period*** (Bailey, 1975). For malaria this is approximately 5-9 days. The ***serial interval*** refers to the time from the appearance of symptoms in the first case until another person is infected and shows symptoms. (Bailey, 1975) With malaria this period will be affected by both human and mosquito population characteristics.

An advanced mathematical model could theoretically use a joint probability distribution for the lengths of the latent and infectious periods, and for the time until symptoms emerge. Due to the complexity of this, many researchers have made various simplifying assumptions while still managing to retain the most important features of the transmission mechanism in question. (Bailey, 1975)

The probability of an infective person actually transmitting to a susceptible depends on various factors such as the organism's virulence and the immunity of the susceptible. In chain-binomial models this is encompassed in a single parameter called 'adequate contact' that represents the chance of transmission between 2 individuals if one is infected and the other is susceptible. (Bailey, 1975) With malaria this is usually represented by distinguishing between infected and infectious individuals with some process linking the two.

Usually it is assumed that susceptibles and infectives mix together homogeneously. However social behaviour in big populations suggests otherwise and this limitation should be considered when appropriate. (Bailey, 1975) Recent work by Heesterbeek (2002) has considered heterogeneous mixing. For malaria homogeneous mixing would mean that the populations of humans and mosquitoes mix randomly i.e. that everyone has the same chance of being bitten regardless of age, location, vocation etc.

Note that the two terms *parasitaemia* and *gametocytaemia* are used throughout this dissertation to indicate the prevalence of either parasites of gametocytes.

1.2.3 Major insights from epidemic theory

Arguably the most important quantity in the study of epidemics, especially when the possible effects of interventions are being evaluated, is what has become known as 'the basic reproductive rate' and is notated as R_0. It can be defined as the expected number of secondary cases that one case can produce if introduced to a wholly susceptible population. (Heesterbeek, 2002)

R_0 has its roots in demography where it is used as a measure of the expected number of offspring born to one female during her lifetime and was formulated by Dublin & Lotka in 1925 (Heesterbeek, 2002). In epidemiology R_0 was originally formulated by Ross in terms of a critical mosquito population density i.e. there is a density of mosquitoes below which malaria will not be sustainable. Therefore the concept of R_0 in epidemiology began as a critical density rather than as a reproduction threshold concept. (Heesterbeek, 2002)

Lotka's work in the demography discipline led to the reformulation of this critical density in terms of reproduction potential. In epidemiology this reformulation was especially powerful as it is related to the necessary effect of an intervention in order to control or even eliminate infection from a population. However it took many years for this concept to move from demography to epidemiology. (Heesterbeek, 2002)

Kermack & McKendrick generalised Ross's critical density concept and developed their 'threshold theorem' in 1927 (Heesterbeek, 2002). This theorem quantifies the probability distribution for the final size of an epidemic in terms of a parameter μ that equals the average number of susceptibles infected by an infective person (Becker, 1989). This theorem postulates that the introduction of infective cases into a susceptible community would only lead to an epidemic if the density of susceptibles was above some critical level. The epidemic would then reduce the number of susceptibles to a level below the threshold as far as it was originally above the same threshold (Bailey, 1975).

The most practical outcome of the threshold theorem is that an outbreak will be minor if $\mu < 1$ and conversely the probability of an outbreak being ma-

jor will be positive if $\mu > 1$. This conclusion has been shown to be robust to the assumptions. One can therefore attempt to prevent a major outbreak by lowering the value of μ through immunisation of a suitable proportion of the population. (Becker, 1989)

The standard threshold theorem applies to a closed population that is initially free of infection. Despite this, the theorem can still explain certain aspects of the spread of disease when the disease is endemic. Endemicity is achieved when the susceptible population is replenished through mechanisms such as births, migration, and loss of immunity. A 'recurrent epidemic' model can be applied to an endemic situation as follows: after a large incidence of disease, the susceptible population will be reduced to a level below the threshold. Another outbreak will only occur once this susceptible population has been replenished back above the level of the threshold. (Becker, 1989)

In 1952 MacDonald produced a reproduction formulation for malaria in terms of the main factors that were identified by Ross. He also coined the term 'basic reproduction rate' for R_0 (Heesterbeek, 2002). The most critical insights into the entomological dynamics of malaria transmission come from the classic formula for R_0 (Smith and McKenzie, 2004).

In 1975 Dietz was the one to clearly define R_0 as "The quantity R is called the reproduction rate, since it represents the number of secondary cases that one case can produce if introduced to a susceptible population." Dietz went on to show that the critical threshold condition translates into R > 1 i.e. an infection cannot be sustained if R_0 is less than one. He also managed to show the crucial step that R_0 could easily be related to data and hence estimated. Thus the primary value of R_0 lies in its threshold value of one. (Heesterbeek, 2002)

At a conference in 1982 Bob May and Roy Anderson widely advocated and promoted the application and value of R_0. Over the last 10-15 years R_0 has been used extensively in the mathematical modeling of infectious diseases and has been applied to increasingly complex and more realistic scenarios. (Heesterbeek, 2002)

1.2.4 Methodological aspects

One of the first choices to be made is usually between a stochastic or a deterministic model. Most modern approaches begin with a deterministic model which might subsequently be extended into a stochastic model. Stochastic treatment is important when dealing with small groups where large variation is likely to occur. (Becker, 1989; Bailey, 1975)

On the other hand, as Thompson (1989) states, a deterministic model can

accommodate most relevant aspects when one is dealing with large populations where the effect of statistical fluctuations is likely to be smaller. In such cases very little is gained from a stochastic model. Becker (1989) calls deterministic models useful for the enrichment of general epidemic theory. Purely deterministic work in the areas of malaria and tuberculosis have been valuable (Bailey, 1975).

Bailey (1975) does warn that while using a deterministic model for large populations of susceptibles and infectives that mix homogeneously is usually satisfactory, it might still be difficult to account for certain large-scale phenomenon without including stochastic processes.

One of the drawbacks of stochastic models is that they lead to mathematical equations that are very difficult to solve explicitly. However computing advances have enabled the complex estimation of parameters, often through iterative procedures, to be easily accomplished (Becker, 1989). Furthermore, modern advances in computer power now allow us to use simulation thereby moving from a set of differential equations that must be solved to using a set of difference equations. Results can then be produced from a computer performing a large number of simple, repetitive operations. (Thompson, 1989, 2000)

1.3 Characteristics of Malaria

This section begins with a general introduction about the burden of malaria and its epidemiology before moving on to an examination of the disease's main characteristics and how these various factors impact on the dynamics of transmission.

1.3.1 Disease burden

Malaria is a potentially life-threatening parasitic disease that is transmitted to humans by mosquitoes, which Barnes and White (2005) and many others call"...the most important parasitic disease of man". The parasite has a complex life cycle and the various stages of this cycle enable the parasite to evade the immune system and infect the liver and red blood cells. Death can be caused through the infection and destruction of red blood cells, and by clogging the capillaries that carry blood to vital organs. (Roll Back Malaria, 2006; Barnes and White, 2005; Diebner et al., 2000; MacDonald, 1957).

The 2005 World Malaria Report estimates that over 3 billion people in over 100 countries are at risk of malaria. Malaria causes somewhere between 300 and 600 million acute illnesses and at least one million deaths annually. (Roll Back Malaria, 2006; Sachs and Malaney, 2002; Snow et al., 2005) Sachs and Malaney (2002) describe the enormous economic burden that malarious places incur and

suggest that there could be as much as a five-fold difference in average GDP between malarious and non-malarious countries.

Africa carries most of the burden of disease with young children being most severely affected and *Plasmodium falciparum* being a leading cause of mortality on the continent (Roll Back Malaria, 2006; Snow et al., 1999b, 2005; Korenromp et al., 2003). In 1995 there were approximately one million deaths and 200 million clinical attacks attributed to malaria in areas of stable transmission in sub-Saharan Africa and about 2000 deaths and 200,000 clinical episodes in areas of unstable malaria in southern africa (Snow et al., 1999a). However it is difficult to get accurate estimates of malaria mortality due to the fact that often such deaths occur in communities without any system of death certification. In many parts of Africa, death in children occurs at home without contact with any formal health service. (Greenwood, 1997; Korenromp et al., 2003). In Eastern and Southern Africa, the past 2 decades has seen a dramatic increase in the malaria mortality rate whereas previous to this, deaths due to malaria were on the decline. This fairly recent upsurge has coincided with a decrease in general mortality among children and increased resistance to widely used antimalarials (Snow et al., 2001; Korenromp et al., 2003).

The characteristics of malaria vary depending, among other things, on the level of endemicity in an area. In non-endemic areas death is usually attributed to cerebral malaria, while death in endemic areas is usually caused by severe anaemia (Greenwood, 1997). The clinical manifestation of malaria is also different in adults and in children, with young children being more likely to develop anaemia and less likely to develop cerebral malaria (independent of acquired immunity)(Luxemburger et al., 1997).

Children and pregnant women are most vulnerable in areas of high intensity. The entire population is at risk in areas of low or unstable transmission, though pregnant women are still especially at high risk and Luxemburger et al. (1997) found that pregnant women are three times as likely as non-pregnant women to develop severe malaria in an unstable transmission area.

Studies into the possible interaction between HIV and malaria, while not being conclusive, have suggested that HIV could impair the immune response to malaria and hence exacerbate the devastating effect of malaria in populations with high HIV rates (Craig et al., 2004; Whitworth et al., 2000; Verhoeff et al., 1999). Craig et al. (2004) also suggest that HIV infection could be related to the spread of drug resistance and so be at least partly responsible for the increasing burden of disease in Africa.

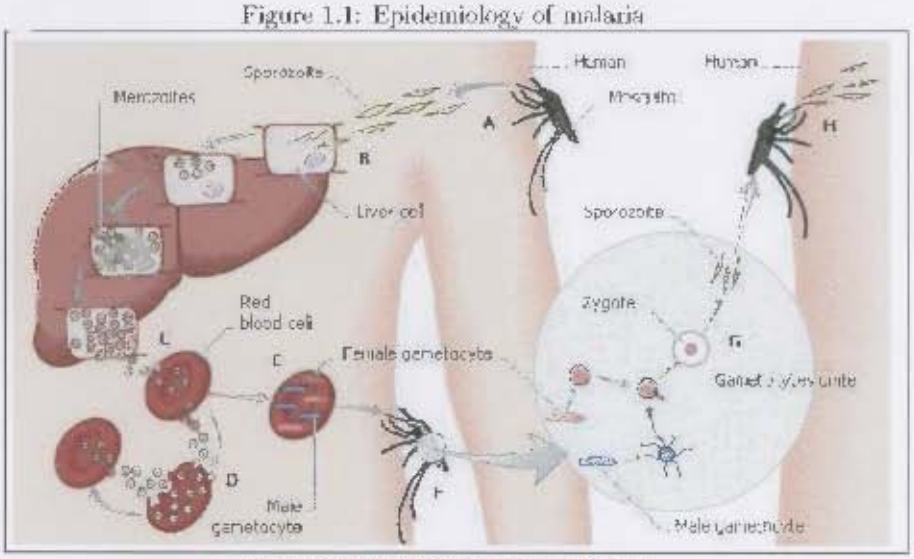
1.3.2 Epidemiology of malaria

There are four distinct species of parasite that cause malaria in humans with the species *Plasmodium falciparum* being the most deadly (Aron, 1988). Fur-

thermore, in Africa, *Plasmodium falciparum* is by far the most widespread of these species. The four *Plasmodium* species are: *falciparum*, *malariae*, *ovale*, and *vivax*. (Roll Back Malaria, 2006)

The life-cycle of the parasite is depicted in figure 1.1 below and can be summarised as follows (Kleinschmidt, 2001; Bailey, 1975; Barnes and White, 2005; MacDonald, 1957; White, 2004; Simpson et al., 2002):

Figure 1.1: Epidemiology of malaria



Inside a human:

1. a person is bitten by an infected female mosquito which results in sporozoites entering the human body.

2. asexual parasites multiply in the liver (incubation period of 5-9 days) in what is called the 'pre-erythrocytic' stage before invading the red blood cells as merozoites.

3. a repetitive cycle begins where these merozoites develop into trophozoites (this is what is visible in blood smears) and divide via the process of schizogony to produce more merozoites which invade non-infected red blood cells.

4. some of the merozoites develop into new trophozoites while others develop into male or female gametocytes (after 2-3 cycles). The actual dynamics that trigger the switch from the asexual stage to gametocytes are unclear (Barnes and White, 2005; Robert et al., 2000; Drakeley et al., 1999; McKenzie and Bossert, 1998).

5. initially the gametocytes resemble trophozoites and are sequestered.

6. They then mature in the small capillaries and venules before being released into circulation.

7. Mature gametocytes appear in the bloodstream 10-12 days after clinical symptoms. Therefore it is the post-clinical period that is most important for the analysis of gametocytes (Jeffery and Eyles, 1955).

A mosquito then gets infected if she feeds on a person with mature male and female gametocytes in their bloodstream. Zygotes are formed which change into ookinetes and then oocysts that are found in the mid-gut wall of the mosquito. Sporozoites are formed within the oocysts (this rate of development is temperature dependent) and invade the mosquito's salivary glands. This phase of reproduction is called 'sporogony' and the mosquito has to survive through this phase in order to transmit an infection.

The initial infection caused by infective sporozoites is called the 'primary phase'. If the infection is treated successfully then this is the only phase that would occur. If the treatment fails, the asexual parasites reappear after a period of time and this is called the 'recrudescent phase'. The recrudescent infection would then either respond to treatment and be eliminated, or else it could recrudesce again. Finally the 3rd stage is called the 'sexual stage' and occurs in both primary and recrudescent infections. (Pongtavornpinyo, 2006)

Asexual parasites, in the absence of death or treatment, typically display a wave-like periodicity where the risk of severe disease is concentrated during the first wave (Dietz et al., 2006). The distribution of asexual parasites over time appears to oscillate after an initial exponential rise. The reason for this is a process known as sequestration whereby approximately halfway through the asexual cycle the infected red blood cells stick to the capillary and venular walls and therefore cannot be detected on a peripheral blood smear. They are later released into circulation. However not all patients exhibit this oscillation and this could be explained by the degree of 'synchronicity' of the infection. The age of the asexual parasites is uniformly distributed in an asynchronous infection and are of a similar age in a synchronous infection. Large amplitudes have been observed for highly synchronous infections while asynchronous infections have amplitudes that approach zero.(Simpson et al., 2002; White et al., 1992)

1.3.3 Transmission

While transmission of malaria is dependent on various factors such as vector characteristics, host susceptibility and climatic conditions, it is host infectivity that plays a crucial role in the spread of an individual infection (Draper, 1953; Diebner et al., 2000; Killeen et al., 2006). Human infectivity is the probability

of a mosquito becoming infected after biting an infected person. Human susceptibility is the likelihood of getting an infection after receiving an infectious bite and unfortunately cannot be measured directly.

Following from the epidemiology of malaria presented above, the asexual parasite needs to progress through its life cycle and develop gametocytes in order for an infection to be transmitted to the vector. There are various factors that have been identified as having an influence on host infectivity including the density of gametocytes, their level of maturity, the proportion of male and female gametocytes, the susceptibility of the mosquito, the size of the blood meal, and the atmospheric conditions. While there is therefore not a single factor that can be measured to represent infectivity, one could expect to find a correlation between the density of gametocytes in a carrier and their infectivity to mosquitoes. (Draper, 1953)

Draper (1953) mentions various studies in different countries in the 30's and 40's that detected such a correlation when densities were categorised but the seminal work of Jeffery and Eyles (1955) appears to be more commonly referenced and widely used. They studied gametocytes when neurosyphilis patients were treated with malaria, and showed a strong association between gametocyte density and the probability of infecting a mosquito.

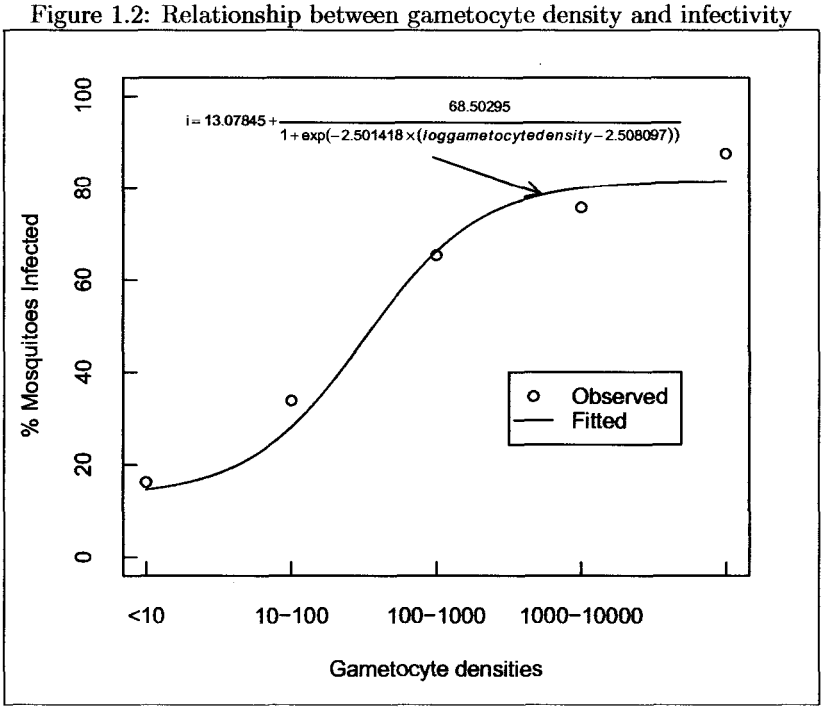
Using the data from Jeffery and Eyles (1955), Barnes and White (2005) showed that the relationship between gametocyte density and infectivity can be described by a log-sigmoid curve as depicted in figure 1.2 below.

More recently, many studies into infectivity and gametocytes have confirmed this association between gametocyte carriage and infectivity (Drakeley et al., 1999, 2004; Targett et al., 2001; Mulder et al., 1994; Tchuinkam et al., 1993), while some have raised doubts about this relationship (Haji et al., 1996).

Note that even those studies where a positive association was found reported a few anomalous cases i.e. where carriers of high gametocyte densities did not infect any mosquitoes and vice versa. The lowest gametocyte density at which transmission can still occur is close to the detectable limit of 8 gametocytes per micro litre (μL)(Barnes and White, 2005). Immunity could also affect the infectivity of gametocytes (Targett et al., 2001). These factors could explain some of the anomalous cases where apparently gametocyte-negative blood caused an infection, or where a carrier with a high density of gametocytes was not infective. Note that for population modeling these anomalous cases should balance each other out i.e. there will be some carriers with low densities that cause infection contrary to expectations but then there will also be those with high densities that are actually not infective.

There have been contradictory findings on the effects of asexual parasite density on gametocyte carriage, which confirms the fact that this relationship cannot be oversimplified. While one would expect a higher asexual parasite

Figure 1.2: Relationship between gametocyte density and infectivity



load to be associated with a higher sexual parasite load, there are also other factors that could be associated with low asexual densities such as immunity, treatment or a long duration of infection. As some of these factors would also increase gametocyte carriage, the relationship in question is often confounded in a clinical setting. (Price et al., 1999; Barnes and White, 2005)

The prevalence of malaria can be high and stable even in areas of lower intensity transmission. Gu et al. (2003b) found that the prevalence of parasitaemia was significantly related to the level of exposure (measured by the Entomological Inoculation Rate or EIR) from 10-11 months previously. This could suggest that gametocytes are carried for a long period and could explain why malaria is maintained over a seasonal or intermittent transmission period.

It is clear that the whole issue of gametocyte carriage and infectivity of humans to mosquitoes is complex and not particularly well understood. However it is also clear that gametocytaemia is critical to the transmission of malaria and spread of drug resistance and definitely warrants further research.

1.3.4 Drug resistance

The development of resistance to anti-malarial drugs by the parasite is generally acknowledged to be the biggest threat to controlling malaria. Unfortunately cheap anti-malarials with oral regimens that are both safe and effective (like chloroquine and sulfadoxine pyrimethamine (SP)) have been affected by large-scale resistance. This has had a detrimental effect on malaria control and consequently has increased the burden of disease (Trape, 2001; Targett et al., 2001; White, 2004; Roper et al., 2003; Nosten et al., 2000; Snow et al., 2001; Babiker et al., 2005).

In South East Asia chloroquine was replaced with SP and then mefloquine but resistance to mefloquine was first noted just 6 years after it was deployed in Thailand (Nosten et al., 2000). Halofantrine shares cross-resistance with mefloquine and its efficacy has also declined (Price et al. (1996); Pongtavornpinyo (2006).

There is now general acceptance that treatment should consist of a combination of 2 or more drugs (as opposed to a single anti-malarial drug) because the probability of a parasite mutating successfully against 2 different drugs with different mechanisms of action is dramatically reduced. Furthermore *Plasmodium falciparum* has developed resistance to all classes of anti-malarials with the possible exception of artemisinin derivatives, where clinically significant resistance has not been identified (Barnes and White, 2005; White, 2004; Pongtavornpinyo, 2006). Therefore optimal treatment should consist of 2 drugs with an artemisinin derivative as one of the drugs in the combination (Barnes and White, 2005; White, 2004; Nosten et al., 2000; Muheki et al., 2004; Targett et al., 2001; WHO Press, 2006).

Many studies have confirmed that a human's infectivity to mosquitoes is greater in a recrudescent or in a resistant infection compared to a primary or a sensitive infection (Price et al., 1996, 1999; Barnes and White, 2005; Robert et al., 2000; Hogh et al., 1998). In fact the rapid spread of resistance to SP is attributed to the high post-treatment prevalence and density of gametocytes carrying the resistant strain (Barnes and White, 2005). The spread of resistance could therefore be reduced if treatment was with a drug that acted on gametocytes (Targett et al., 2001; Barnes and White, 2005).

1.3.5 Immunity

In 1900 Robert Koch was the first person to notice the fact that the frequency and density of asexual parasites was lower with increasing age among people in Java, an endemic malaria area in Indonesia (Baird, 1998; Pongtavornpinyo, 2006). Surveys conducted in the 50's, 60's and 70's showed conclusively that in areas of high transmission intensity the prevalence of malaria declines with age

(Greenwood, 1997; Koella, 1991; Aron, 1988). Areas of high intensity transmission exhibit a malaria incidence peak in early childhood, at moderate intensity this peak occurs at a slightly older age, and in low transmission settings the risk of infection remains constant across ages (Snow and Marsh, 1998).

The conventional view is that the gradual acquisition of partial immunity is a consequence of repeated infections over time though Baird et al. (1991) makes a compelling case for immunity to be caused by *recent* exposure history together with intrinsic features of the host's immune system that are dependent on the natural ageing process. Interrupted exposure can result in the loss of immunity (Pongtavornpinyo, 2006; Koella, 1991; Aron and May, 1982; Aron, 1988).

Various studies have concluded that there is a short period of maternal immunity that is conveyed to an infant. A small number of infective bites in early childhood can be sufficient to acquire some immunity. (Gupta et al., 1999; Snow et al., 1998)

Since there is no direct measure of immunity, various clinical outcomes that are stratified by age (asexual parasite density, failure rates, severe malaria rates) have traditionally been used as a proxy. Age-stratified asexual parasite density and prevalence have become the standard proxies of immunity. (Pongtavornpinyo, 2006; Aron and May, 1982)

Though immunity does not confer complete and perfect protection, it leads to a reduction in:

- susceptibility i.e. the probability of developing an infection after getting bitten by an infective mosquito (Dietz, 1988)
- the asexual parasite density (and hence severity of disease) in infected patients (Dietz et al., 1974; White, 2002; Baird et al., 1991; Dietz, 1988)
- the treatment failure rate (Luxemburger et al., 1997; Mayxay et al., 2001; White, 2002)
- the duration of infection (Rogier et al., 1999; Dietz et al., 1974)
- gametocyte infectivity through something called 'Transmission Blocking Immunity (TBI)'. (Graves et al., 1988; Mendis et al., 1987; Mulder et al., 1994). MacDonald (1957), Dietz et al. (1974) and Dietz (1988) also spoke about immunity reducing gametocyte density and thereby infectivity.

Therefore malaria in an immune person often results in an asymptomatic infection due to low levels of parasitaemia. However as low asexual parasite densities can still produce gametocytes and successfully transmit an infection, asymptomatic carriers play an important part in malaria transmission in endemic scenarios where acquired immunity exists (Jeffery and Eyles, 1955; Drakeley et al., 1999; Kleinschmidt, 2001; Barnes and White, 2005).

With drug pressure, if immunity was ignored one would expect an exponential increase in the proportion of cases due to the increase in resistant parasites. In fact this phenomenon is observed in low intensity areas during early epidemics (Pongtavornpinyo, 2006; Koella, 1991). Despite its importance, immunity was only introduced into mathematical models of malaria fairly recently (Aron and May, 1982).

1.3.6 Malaria control and the effect on gametocytes

Arguably the greatest revelation to emerge from the early mathematical models of malaria is the fact that malaria can be eradicated without eliminating transmission completely. Transmission just needs to be reduced below some critical value and the disease will die out. The two main methods of control focus on the vector and on early asexual parasite detection and treatment.

Often treatment is focused on the individual symptomatic patient by eliminating the asexual stages of the parasite. While drugs that act only on asexual stages will prevent new generations of gametocytes being formed (Mulder et al., 1994), it is necessary to also reduce the carriage of existing gametocytes in order to limit malaria transmission and the spread of resistance, otherwise an apparent therapeutic success can lead to resistant genes being passed to the vector (Barnes and White, 2005; Targett et al., 2001; Price et al., 1999).

One goal of malaria treatment selection is to prolong the useful life of antimalarial drugs and therefore any change in treatment policy needs to consider the possibility of resistance developing thereby rendering the new drug obsolete (Pongtavornpinyo, 2006; Targett et al., 2001). The transmission of malaria in areas with low endemicity is likely to stem from symptomatic patients with low levels of acquired immunity that seek treatment. Conversely, transmission in endemic areas also involves asymptomatic parasitaemia from untreated people and hence less drug pressure (White, 2004). In the first case, the fact that a large amount of people are treated confers a survival advantage to the parasites that have developed resistance to the particular drug. This happens less in the latter case. The issues of what drug to switch to and when to make the switch are critically important and depend on the epidemiology in each particular area. (Pongtavornpinyo, 2006)

Different drugs affect different phases of gametocytogenesis. Mature gametocytes are not sensitive to many schizontocidal drugs, including chloroquine and pyrimethamine/sulfadoxine (Price et al., 1999; Hogh et al., 1998; Robert et al., 2000). It is only the 8-aminoquinoline class of drugs (e.g. primaquine) that are effective against mature gametocytes (Robert et al., 2000).

Treatment with an artesunate combination was examined in several studies and resulted in significantly lower prevalence and density of gametocytes and

hence the authors conclude that artesunate does act on the sexual stage of the parasite (Targett et al., 2001; Price et al., 1996). It is thought that this action stems from 2 different mechanisms: firstly by acting on young gametocytes and secondly as a consequence of rapidly decreasing asexual parasite loads. Chloroquine only seemed to affect development of very young gametocytes (Targett et al., 2001).

SP is associated with the highest post-treatment gametocytaemia (Barnes and White, 2005). Robert et al. (2000) found that post-treatment gametocytaemia (in terms of both prevalence and density) was found to be higher for SP patients compared to patients on chloroquine, while Targett et al. (2001) found that treatment with SP alone compared to chloroquine or combination therapy, resulted in a dramatic surge of gametocytes although these were less infective. The relatively lower infectivity of the gametocytes could be due to the fact that the drug causes the release of sequestered gametocytes that are not yet infectious. However the higher gametocyte density more than offsets the lower risk of transmission.

While the effect of immunity on transmission may be relatively unimportant for an eradication theory it is critical for a theory of control whose endpoint is a new equilibrium between host and parasite populations (Dietz et al., 1974; Bailey, 1975). Immunity can also make failing drugs appear to be effective (White, 2002). In some scenarios control measures can even end up having counter-intuitive effects by reducing the level of immunity in a population (Snow and Marsh, 1998).

Gu et al. (2003a) used an individual-based model to investigate control strategies via simulation. They found that extinction was unlikely in high transmission areas but quite feasible in low transmission areas. The probability of eradicating malaria locally was very sensitive to migration by infected people and they found that even a small amount of migration could prevent local extinction with the introduction of even a single case having the potential to spark an epidemic. This has particular relevance in light of the fact that the majority of malaria models assume closed populations.

1.4 Previous Malaria Models

The purpose of this section is to review some of the major malaria models that have been previously developed. I begin with a fairly in-depth look at the Ross and Ross-MacDonald models as these form the foundation for future malaria modeling. I then briefly examine some of the major extensions that have been implemented over the past few decades before finally presenting the concept of a gametocyte-driven model. Note that the objective of this section is not to present a detailed mathematical view of the models but rather an overview with

specific focus on how host infectivity is handled.

1.4.1 The classics

Ross

In 1898 Ronald Ross discovered that malaria was spread by mosquitoes. He was also the first to try and develop a theory about the underlying mechanism in the spread of infection by using *a priori* assumptions (Heesterbeek, 2002; Bailey, 1975; Fine, 1975), and developed the first deterministic model of malaria transmission in the early 1900's.

Bailey (1975) presents a description of the foundations of Ross's classic model (and Ross (1911) was also consulted). The following parameters were used for the human population:

- t: represents the time variable
- n: total population size at a given time
- y: total number of infected individuals
- f: proportion of infected individuals that are infective
- γ: recovery-rate
- μ: birth-rate
- v: death-rate

The same set of parameters with primes i.e. y' are applied to the mosquito population along with the single parameter b' that represents the man-biting rate.

According to Ross's classic model, in time Δt, y' infected mosquitoes make $b'f'y'\Delta t$ infectious bites and only a proportion of these bites ($\frac{n-y}{n}$) are on susceptible humans. The analogy is similar for the mosquitoes though note that the denominator remains in terms of the human population n. Heuristically, $(n'-y')$ susceptible mosquitoes make $(n'-y') \times b'$ bites and a proportion of these bites are on humans that are both infected and infective $\frac{yf}{n}$.

This leads to the following differential equations that describe the rate at which the infected populations of humans (y) and mosquitoes (y') grow:

$$\frac{dy}{dt} = \frac{b'f'y'(n-y)}{n} - (\gamma + v)y \tag{1.1}$$

$$\frac{dy'}{dt} = \frac{b'fy(n'-y')}{n} - (\gamma' + v')y'. \tag{1.2}$$

Furthermore Ross spoke of the fact that the human recovery rate was much faster than the human death rate and the opposite is true for mosquitoes and hence v was negligible relative to γ and γ' was negligible relative to v'. Lastly it is assumed that the birth and death rates are equal and hence v' can be replaced by μ'.

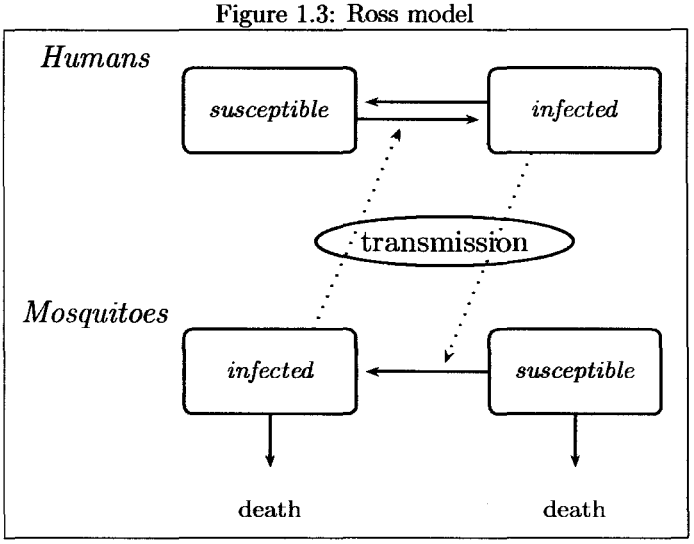
Equations 1.1 & 1.2 can then be simplified to:

$$\frac{dy}{dt} = \frac{b'f'y'(n-y)}{n} - \gamma y \tag{1.3}$$

$$\frac{dy'}{dt} = \frac{b'fy(n'-y')}{n} - \mu'y' \tag{1.4}$$

This simple model is depicted in figure 1.3 below:

Figure 1.3: Ross model



Humans move back and forth between a susceptible and an infected state whereas mosquitoes move from a susceptible to an infected state and are removed from the system by deaths.

Equations 1.3 & 1.4 can be rewritten in terms of $m=\frac{y}{n}$ the malaria rate in man, $u=\frac{y'}{n}$ the density of infected mosquitoes per human, and $a=\frac{n'}{n}$ the overall mosquito density per human:

$$\frac{dm}{dt} = b'f'u(1-m) - \gamma m \tag{1.5}$$

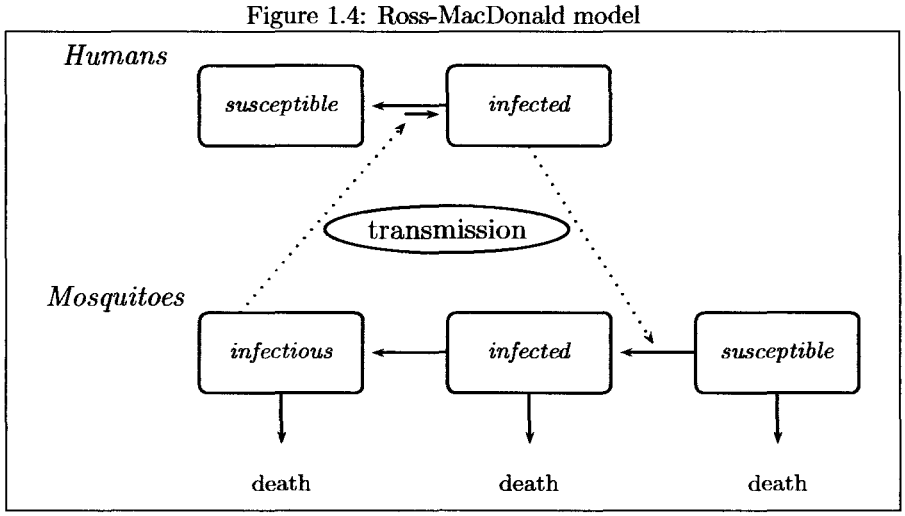
$$\frac{du}{dt} = b'fm(a-u) - \mu'u \tag{1.6}$$

Ross then derived his critical density 'mosquito theorem' from these equations. Essentially he found that there is a critical density, in terms of the ratio of mosquitoes to man, below which malaria could not sustain itself. This insight had important consequences for control since it led to the realisation that malaria could be controlled without having to eradicate the entire mosquito population (Heesterbeek, 2002; Bailey, 1975; Fine, 1975; Molineaux, 1985).

MacDonald

MacDonald extended Ross's basic model into what is know as the 'Ross-MacDonald' model (MacDonald, 1957; Koella, 1991).

Figure 1.4: Ross-MacDonald model



Examining figure 1.4, it can be seen that one addition is the incorporation of the time period for sporozoites to form within the mosquito i.e. mosquitoes have to survive through this period in order to be capable of transmitting infection. Without this the basic model predicts unrealistically high prevalences of infected mosquitoes (Aron and May, 1982). The other major extension tackled by MacDonald was an attempt to include super-infection.

Super-infection occurs commonly in areas of high intensity and refers to the case where a person carries multiple infections. MacDonald (1950) found that analysing data with Ross's model led to fairly well-fitting curves but unrealistic recovery rates. He attributed this to the fact that Ross had ignored super-infection and went on to incorporate reinfection in his model (Bailey, 1975; Dietz et al., 1974).

As described in section 1.2.3, the basic reproductive rate or R_0 is a key

parameter in the analysis of infectious diseases. One of the most important results to emerge from the Ross-MacDonald model was the derivation of R_0 for malaria (Koella, 1991; Bailey, 1975; Aron and May, 1982). The reproductive rate describes the number of secondary cases that arise from a single case in an otherwise uninfected population:

$$R_0 = \frac{ma^2 b_1 b_2 \exp^{-\mu T}}{r\mu} \tag{1.7}$$

where m is the number of mosquitoes per human host, a is the biting rate, b_1 is the infectivity of humans to mosquitoes, b_2 is the susceptibility of humans, μ is the mortality of adult mosquitoes, T is the incubation period of parasites within the mosquito, and r is the recovery rate of infected humans. According to Dietz (1988) it is typical to assume values of one for both b_1 and b_2 which can lead to R_0 being severely overestimated.

This equation makes intuitive sense since transmission is increased by high densities of mosquitoes that bite infective humans frequently and transmit the infection to susceptible humans, whereas transmission is hindered by a quick recovery rate and by a high mortality rate of the vector. As two bites are required for an infection to be transmitted, a enters the equation twice. The $\exp^{-\mu T}$ term represents the proportion of mosquitoes that survive from the time of being infected until sporozoites are developed in their salivary glands. (Koella, 1991)

An alternative notation for R_0 presented in Aron and May (1982) is:

$$R_0 = \frac{ma^2 b \exp^{-\mu T}}{r\mu} \tag{1.8}$$

where b is now the proportion of bites by infected mosquitoes that result in infection.

It is possible to then write the equations for the prevalences of infected humans (y) and for infectious mosquitoes (w) in terms of R_0. This shows that R_0 needs to be at least one for malaria to be maintained in a population. When the reproductive number is near one, a small increase in R_0 lead to a large increase in prevalence. On the other hand, if R_0 is large then even large reductions in R_0 lead to nominal reductions in prevalence. Lastly, for a very high R_0 the model predicts that virtually the entire population is infected. This obviously is due to the omission of acquired immunity:

$$\hat{y} = \frac{R_0 - 1}{R_0 - \frac{a}{\mu}} \tag{1.9}$$

$$\hat{w} = \frac{R_0 - 1}{R_0} \frac{\frac{a}{\mu}}{1 + \frac{a}{\mu}} e^{-\mu T} \tag{1.10}$$

(Koella, 1991)

MacDonald used a more sophisticated formulation to represent the probability and duration of survival of the mosquito after the extrinsic incubation period:

$$\frac{p^n}{-lnp}$$

that assumes a constant daily probability of survival p (MacDonald, 1957). This was an important difference as it allowed MacDonald to assess potential effects on this daily survival rate through interventions like residual insecticides. (Fine, 1975)

In fact, sensitivity analysis revealed that the largest reduction in R_0 is found when adult mosquito mortality is increased and this leads to the important conclusion that imagicides are a more effective intervention than larvicides (Koella, 1991; Aron and May, 1982; MacDonald, 1957). Smith and McKenzie (2004) found that increasing adult mosquito mortality leads to an even larger effect than that found by MacDonald.

MacDonald (1957) also derived the formula for the inoculation rate (which is similar to the first term in equations 1.1 and 1.3 and refers to the rate at which humans are infected) h as:

$$h = mabs \tag{1.11}$$

where: m is the mosquito density per human, a is the average number of humans bitten per day by any one mosquito, s is the proportion of mosquitoes with sporozoites in their glands, and b is the proportion of mosquitoes that have sporozoites in their glands and are actually infectious.

and the following formula for s:

$$s = \frac{p^n ax}{ax - ln(p)} \tag{1.12}$$

The Ross-MacDonald model helps to interpret differences between endemic situations and is useful in predicting the major impacts of control strategies. However it is a relatively simple model that makes a number of assumptions: all newborns are uninfected and susceptible, there is no immunity, there is random biting, and the populations are homogeneous and closed. It is less useful at explaining the dynamics of malaria in a given area or at assisting to design malaria control interventions strategies e.g. where and how often should insecticides be applied. (Koella, 1991)

Dietz

Macdonald's mathematical treatment of super-infection actually assumed that an individual could only recover from one infection at a time and hence considerably overestimated the duration of infection for high inoculation rates (Bailey, 1975; Dietz, 1988). Dietz developed a new formula for the recovery rate in the presence of super-infection according to a Poisson process with inoculations 'arriving' with rate h and an exponentially distributed duration with mean r^{-1} (Dietz et al., 1974).

The average duration of an infection (T) was then derived as follows: the probability of an individual having no infections is $exp^{-\frac{h}{r}}$ and the waiting time for a new inoculation is h^{-1}, which leads to the equation:

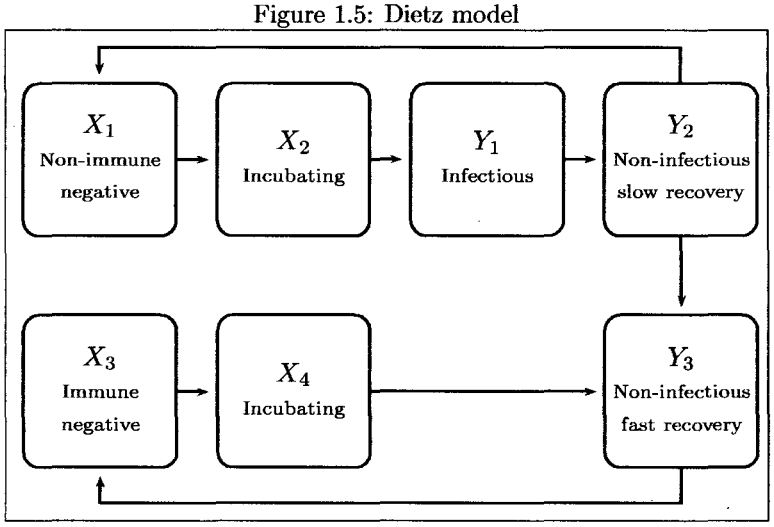
$$exp^{-\frac{h}{r}} = \frac{h^{-1}}{T + h^{-1}} \tag{1.13}$$

and this can be written in terms of T as:

$$T = \frac{[exp^{-\frac{h}{r}-1}]}{h} \tag{1.14}$$

Perhaps more importantly, Dietz also attempted to incorporate the effect of immunity on transmission. He did this by setting up 2 different immunity classes, one that had a slow recovery rate from malaria with all infections being detected, and the other that had a fast recovery rate from malaria and a 70% chance of an infection being detected. Both classes are exposed repeatedly, contract malaria, and recover. Individuals remain within their class except for a fixed rate of transition from the non-infective slow recovery state to the non-infective fast recovery state:

Figure 1.5: Dietz model



The model basically works as follows: newborns enter the class of non-immune negatives. The inoculation rate h moves individuals into the incubating state (i.e. asexual parasites in the liver only) where they spend a fixed length of time before moving into the infective positive state. Infectivity is lost at a constant rate causing individuals to move to the non-infective positive state. From here they either lose their immunity, recover and move back to the non-immune negative state at rate R_1, or they move to another class (immune) of non-infective positives. The immune non-infective positives retain their immunity and move to the class of immune negatives at rate R_2. The recovery rates R_1 and R_2 are derived for a scenario with super-infection. (Dietz et al., 1974; Aron and May, 1982)

Note that figure 1.5 only depicts the change in the human population and does not model the mosquito population explicitly. Rather the inoculation rate h needs to be a function of time that depends on the dynamics of the mosquito population. The relevant vector dynamics are captured in a single variable C(t) called the vectorial capacity that depends on entomological parameters but not the asexual parasite or sporozoite rates. This value represents the number of infections that the vector population distributes per case per day (Garrett-Jones, 1964):

$$C(t) = \frac{m(t)a^2p^n}{-ln(p)} \tag{1.15}$$

where m = vector density, a = man-biting rate p = daily survival probability, and n = the incubation period of the parasite in the mosquito. a p, and m vary with mosquito species and a p, and n can be time-dependent. $-ln(p)^{-1}$ is the expectation of the longevity of a mosquito. MacDonald (1957) assumes that the principal causes of death in mosquitoes are the hazards of daily life and hence the survival rate can be considered a parameter that is independent of the mosquitoes age.

MacDonalds's formula for the inoculation rate given in equation (1.11) requires a measure for b that is not easy to obtain and so a new formula for inoculation rate was invented by Dietz et al. (1974) that used a human susceptibility parameter *(g)*:

$$h(t) = g[1 - exp(-C(t-n)Y_1(t-n))] \tag{1.16}$$

where g is the conditional probability that an infection results given at least one contact has occurred, C is vectorial capacity, and Y_1 is the proportion of the population that are infected and infectious.

The exponent term is the average number of potentially infective contacts that the infective positives on day (t-n) make on day t. The part in the square brackets represents the probability that at least one contact is made assuming

a poisson distribution for the number of contacts. The parameter g is then the conditional probability that an infection occurs, given that at least one contact has been made, also called *susceptibility*. Dietz used a data-driven approach to estimate the g parameter.

1.4.2 Recent developments

Bangkok model

The 'Bangkok model' was developed by Pongtavornpinyo (2006) and attempts to address the fact that other models have oversimplified the complex relationship between transmission intensity, immunity, clinical malaria, and the spread of resistance. It therefore incorporates key host, vector, parasite and drug components and attempts to model antimalarial drug resistance. It specifically aims to explore how much of the benefit of Artemisinin-based Combination Therapy (ACT) is due to delaying resistance versus the short term benefits of affecting a cure and reducing transmission.

A simplified version of R_0 was used here to model the spread of resistance. The relative reproductive rate was used and represented by the ratio of the reproductive rates in drug resistant versus drug sensitive parasites.

Human infectivity is based on the relationship between gametocyte density and the probability of infecting mosquitoes (see figure 1.2 in section 1.3.3). Sexual densities were calculated directly from asexual densities using an estimated switching rate that depends on the type of drug used.

Stochastic models

Gu et al. (2003a) developed an individual-based model of *Plasmodium falciparum* transmission. An object-oriented approach was used with two objects (namely humans and mosquitoes) that both exhibited individual variability for parameters such as recovery and survival rates. Both humans and mosquitoes had three possible states namely susceptible, infected and infectious. They assumed that infected people that carry the infection beyond the intrinsic incubation period of 15 days are infective to mosquitoes, and that mosquitoes surviving the extrinsic incubation period of 10 days likewise become infectious to humans.

In August 2006, the American Journal of Tropical Medicine and Hygiene published a supplement on 'Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of *Plasmodium falciparum* malaria'. By this time most of the research for this dissertation had been completed. The main thrust of the supplement is a stochastic simulation model that gives stochastic predictions of asexual parasite densities based on

the duration of the infection and on immunity status (Smith et al., 2006), and the model actually consists of several sub-models incorporating various aspects of the disease dynamics.

The selection of papers in this supplement illustrates the advances in the use of mathematical models and stochastic simulation to model different features of malaria. In particular, three papers are of interest to this research:

1. Maire et al. (2006) derived models for asexual parasite densities to determine the impact of immunity on these densities. They used a stochastic simulation model.

2. Ross et al. (2006) examined the relationship between infectivity and asexual parasite density, taking into account the time lag caused by the epidemiology of gametocytaemia. They assume that the density of functional female gametocytes is related to the asexual parasite density by: $ln(y_g(i,t)) \sim N(ln(\rho\gamma(i,t)), \sigma_g^2)$. This is based on the fact that the ratio of gametocytes to asexual parasites is lognormally distributed with a geometric mean ρ. Whilst taking both sexual and asexual densities into account in the analysis, they did not directly fit a model to gametocyte densities. There were two reasons given for modeling gametocytes indirectly: due to the lack of understanding of gametocytogenesis, and due to the fact that a model of infectivity as a function of asexual parasites was required.

3. Dietz et al. (2006) fitted a mathematical model to asexual densities and used simulation to explore the effects of vaccination at the individual level. They fitted simple piecewise linear models using least squares to the log transformed observed asexual parasite densities.

While Ross et al. (2006) come closest to looking at gametocyte densities, all of these papers focus on the modeling of asexual parasite densities. In contrast, this dissertation focuses on modeling gametocyte densities over time directly.

Nonlinear mixed effects modeling

Simpson et al. (2002) used nonlinear mixed modeling to model the first wave of asexual parasite densities using the following function:

$$log_{10}y = a + (0.5 \times log_{10}PMR \times t) + c \times sin(\frac{2\pi}{period} \times t + k) \qquad (1.17)$$

The first two terms were derived from the exponential growth of parasitised red blood cells, assuming a constant growth rate of b and with an asexual parasite multiplication rate every two days equal to 10^{2b}, while the last term accounts for the oscillation in the density-time profiles as a result of sequestration.

The objective was to characterise the initial growth dynamics of the infection before the host defence is likely to have a significant effect. Two meaningful parameters namely the Parasite Multiplication Rate (PMR) and the length of the asexual parasite life-cycle (period) were estimated. The other parameters that were estimated were c (amplitude of the sine wave), k (the phase shift in the sine wave) and a (asexual parasite density at day 0).

There was a large amount of variability in the asexual parasite density profiles between patients. This inter-individual variability in $PMR, c, period$ and k were modelled by a log-normal error model, whereas the inter-individual variation in the a parameter was modelled with an additive error model.

The wave-like structure fitted adequately for approximately half the patients with some patients not exhibiting oscillation. This could be related to the synchronicity of the infection where highly synchronous infections exhibited large amplitudes and asynchronous infections had amplitudes that approached zero.

The model was fitted to various subsets of the data, i.e. to only those exhibiting a wave-like pattern in their response and then to the full dataset, and the PMR estimate did not change much for the full dataset. They also fitted a patient group variable with 4 levels based on the observed response pattern as a covariate.

The authors conclude that the fact that the model did not fit well for certain patients does not necessarily imply that the model was mis-specified. This could be caused by discrepancies between the observed asexual parasite count and the actual total asexual parasite burden.

The approach used has direct relevance for this research due to the fact that the same methodology was applied, and patients were grouped according to the observed shape of their density-time profiles. One further relevant observation that the authors make is that an optimal study to model the oscillation in the data would require more frequent time points.

1.4.3 A gametocyte-driven model

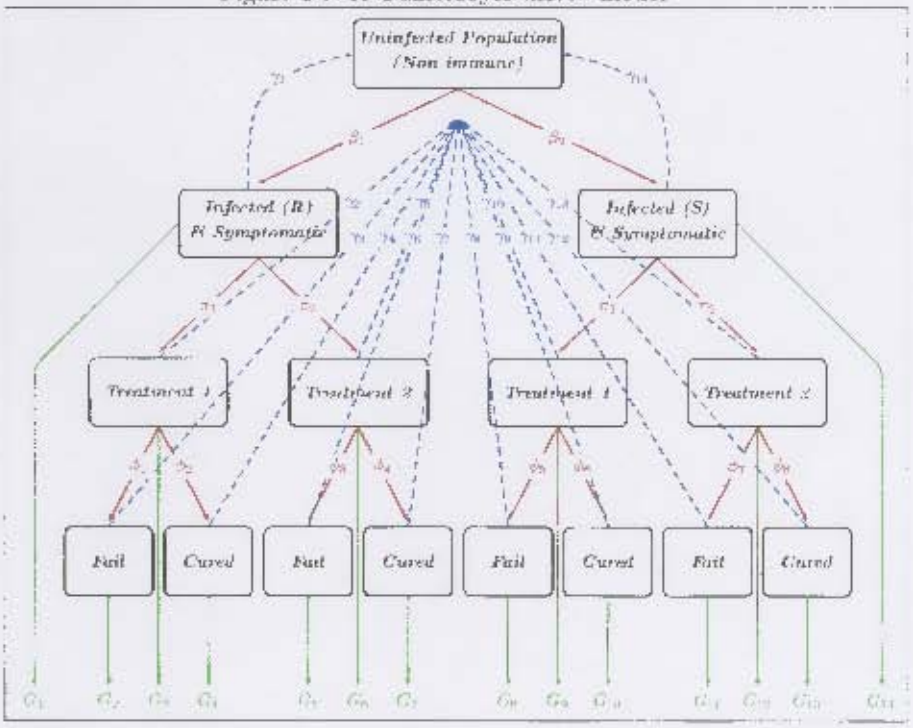
None of the models of malaria transmission reviewed are driven directly by gametocyte densities. The conceptual model presented here illustrates how gametocyte density-time curves could potentially be used to drive a model of malaria transmission.

The basic idea is that one would model gametocytes directly and hence generate a set of different gametocyte curves based on relevant patient profiles. Ideally such a model would be derived from data sets with rich measures of gametocyte densities over time, and could incorporate factors such as age, im-

munity, drug used, treatment response, and level of parasite resistant mutations. These curves could then be used in the model to estimate gametocyte densities and clearance rates, hence leading to an estimate of infectivity in the population.

A conceptual gametocyte-driven model is illustrated in figure 1.6 below (note that fig 1.6 is for a non-immune, closed population where asymptomatic infections are negligible):

Figure 1.6: A Gametocyte-driven model



The parameters for this model would be:

- β_i = infection rates
- π_i = probability of treatment (related to coverage)
- ϕ_i = probability of treatment outcome (success/failure)
- G = gametocyte density
- γ_i = rate of gametocyte clearance

Humans move out of the "uninfected" state according to the infection rates β_i which determine the proportion of humans per unit time that are infected

with either a resistant (R) or a sensitive (S) strain. Treatment is then received with a probability of π_i depending on the particular treatment strategy used, and these rates move people into one of the treatment states. Untreated cases remain infected and hence continue to produce gametocytes (G_1 and G_{14}). Patients are typically in treatment for several weeks and so would stay in the treatment box and be moved into either the "fail" or "cure" state according to the treatment outcome probabilities (ϕ_i). The other main feature of this model is the fact that an infected person only moves back into the "uninfected" state once they have cleared their gametocytes, and these clearance rates are defined by the γ_i parameters. Hence a treatment success could be in the "cured" state and appear to be healthy, yet still carry gametocytes in their blood and hence still contribute to the general reservoir of infectivity.

At the end of each iteration the model would estimate the level of infectivity based on the density of gametocytes circulating in the population (all the various G_i's). This value would be input as a component of the infection rate for the next iteration. As depicted, the model can easily be stratified by treatment and resistance (R=resistant and S=sensitive), thereby allowing malaria transmission and the spread of resistance following alternative treatment strategies to be evaluated.

Note that while humans are the objects that move between the various states in the flow diagram, it is gametocytes that drive the model. Individuals only move back into the uninfected population when they are clear of gametocytes.

1.5 Relevance of Literature Review for this Thesis

This dissertation aims to make a contribution in the realm of malaria modeling and the literature review begins by looking at the modeling of infectious diseases in general in order to broadly contextualise the area that is related to this research.

The section examining characteristics that are specific to malaria is required in order to understand how gametocytes fit into the epidemiology of malaria, their role in the spread of drug resistance, as well as how gametocyte carriage is affected by various factors.

The modeling of gametocyte carriage focuses on a single component of the complex dynamics of malaria transmission and the spread of resistance, namely host infectivity. It is therefore necessary to review the classic models of malaria transmission in order to understand how this component fits into the bigger picture. The brief examination of some of the more recent models of malaria

transmission highlights the fact that despite the obvious importance of gametocyte carriage, very little has been done with regards to modeling this directly.

The literature review ends with a conceptual model that is driven by gametocytes in order to show how gametocyte density-time curves could be utilised. This conceptual model could be stratified by resistance and therefore allow the transmission of malaria as well as the spread of resistance to be modelled, and hence different control strategies could be evaluated.

There is a large amount of variability between individuals in their gametocyte profiles over time, plus the shapes of these profiles are definitely not linear. Nonlinear mixed modeling is therefore the appropriate methodology for modeling gametocyte densities over time as it allows a nonlinear function to be fitted together with individual variability in the parameters of the function.

Chapter 2

Methods - Nonlinear Models for Repeated Measures

The contents of this chapter are taken from the following references: (Pinheiro and Bates, 2002; Davidian and Giltinan, 1998; Ratkowsky, 1990).

2.1 Introduction

Nonlinear models allow for greater flexibility in the description of relationships compared to linear models, but this flexibility comes with a cost due to the increase in computational complexity. One could use a linear polynomial model that is linear in its parameters but this sort of empirical model only holds within the range of the data and the parameters are difficult to interpret contextually. A nonlinear model on the other hand takes a more mechanistic approach i.e. attempts to uncover the underlying mechanism that produced the data, and as such the parameters often have a natural interpretation for example an absorbtion rate. Nonlinear models also are typically more parsimonious then corresponding polynomial models.

Formally, a nonlinear model has at least one of its derivatives with respect to the paramaters being a function of one of those parameters. The parameters cannot be obtained explicitly and hence iterative procedures are used.

Different parameterisations of the same model can produce estimators with different statistical properties even though they produce identical predicted values. Different parameterisations can affect convergence, stability (if paramterised so that parameters are independent), and accuracy of estimation.

It is a fallacy that high correlation amongst the parameters from a nonlinear model may cause convergence difficulties. It can however be indicative of overparameterisation and that can cause issues with convergence.

The relevant theory is presented in two broad sections namely model specification, and inference and estimation. In each section I begin by reviewing the theory for the simpler hierarchical linear model. The reason for starting with this is that the approach used for nonlinear models typically produces an approximation that is in a linear form and consequently the same theory can then be applied. I then look at a nonlinear model for an individual before moving on to the hierarchical nonlinear model.

2.2 Model Specification

With repeated measurements it is important to recognise that there are two levels of variability: random variation within an individual (intra-individual), and random variation between individuals (inter-individual). The models can therefore be specified in two stages.

2.2.1 Hierarchical linear mixed effects models

The subscript i is used to denote the particular individual out of m individuals so that there are a total of $\sum_{i=1}^{m} n_i$ data points for the whole sample. $\boldsymbol{y_i}$ then represents the $(n_i \times 1)$ vector of responses for subject i, $\boldsymbol{\beta}$ represents a $(p \times 1)$ vector of parameters for the p fixed effects, $\boldsymbol{X_i}$ is a $(n_i \times p)$ design matrix for individual i, $\boldsymbol{b_i}$ is a $(k \times 1)$ vector of random effects (i.e. there can be k different random effects), $\boldsymbol{Z_i}$ is a $(n_i \times k)$ design matrix linking the y_i with the random effects, and $\boldsymbol{e_i}$ is the vector of intra-individual random errors. $\boldsymbol{R_i}$ is the inter-individual covariance matrix that depends on i only through the dimension of i.

Stage 1: intra-individual The vector of responses for the ith individual can be written as:

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{e_i} \tag{2.1}$$

Conditional on $\boldsymbol{b_i}$, (2.1) implies that $E(\boldsymbol{y_i} \mid \boldsymbol{b_i}) = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}$ and $Cov(\boldsymbol{y_i} \mid \boldsymbol{b_i}) = \boldsymbol{R_i}$ and hence:

$$\boldsymbol{e_i} \sim N(0, \boldsymbol{R_i}) \tag{2.2}$$

Stage 2: inter-individual Under the assumptions that $\boldsymbol{b_i} \sim N(0, \boldsymbol{D})$ with $\boldsymbol{D}$ being a $(k \times k)$ dispersion matrix for the k random effects, and that the $\boldsymbol{b_i}$ are independent of each other and the $\boldsymbol{e_i}$:

$$E(\boldsymbol{y_i}) = E\Big\{E(\boldsymbol{y_i} \mid \boldsymbol{b_i})\Big\} = \boldsymbol{X_i\beta} \tag{2.3}$$

$$\begin{aligned} Cov(\boldsymbol{y_i}) &= E\Big\{Cov(\boldsymbol{y_i} \mid \boldsymbol{b_i})\Big\} + Cov\Big\{E(\boldsymbol{y_i} \mid \boldsymbol{b_i})\Big\} \\ &= \boldsymbol{R_i} + \boldsymbol{Z_i D Z_i}' \\ &= \boldsymbol{V_i} \end{aligned} \tag{2.4}$$

where $\boldsymbol{V_i}$ typically depends on a few variance parameters notated as the vector of parameters $\boldsymbol{\omega}$.

Therefore, under assumptions of normality and independence for $\boldsymbol{b_i}$ and $\boldsymbol{e_i}$, unconditionally: $\boldsymbol{y_i} \sim N(\boldsymbol{X_i\beta}, \boldsymbol{V_i})$

2.2.2 Nonlinear models for an individual

Data following a nonlinear model often exhibit a mean-variance relationship i.e. the variance of the $\boldsymbol{y_i}$'s is not constant or homogenous across the range of the responses. It is also fairly typical for the observations within a particular subject to be correlated i.e. not independent.

I first begin with the specification for the basic model and then extend it to incorporate heterogeneous variance and intra-individual correlation. The subscript i for individual has been suppressed in this section and the subscript j refers to the jth repeated measurement on an individual.

Basic NL Model

$$y_j = f(\boldsymbol{x_j}, \boldsymbol{\beta}) + e_j \tag{2.5}$$

where for example: $f(\boldsymbol{x_j}, \boldsymbol{\beta})$ could be: $\beta_1 \exp^{-\beta_2 x} + \beta_3 \exp^{-\beta_4 x}$ or any other nonlinear function.

This basic model holds under the usual set of classical assumptions i.e. that the $e_j \sim i.i.d.\ N(0, \sigma^2)$. The assumption of normality forms the basis for the standard approach to inference. Under these assumptions the y_j's are independently, normally distributed with:

$$E(y_j) = f(\boldsymbol{x_j}; \boldsymbol{\beta})\ ; Var(y_j) = \sigma^2 \tag{2.6}$$

As alluded to above, the assumptions of independence and of common variance are frequently violated in the nonlinear paradigm.

Intra-individual variance heterogeneity One way to account for non-constant variance is to model the response variance in a similar way to how one models the mean response: eg $Var(y_j) = \sigma^2 (f(\boldsymbol{x_j}, \boldsymbol{\beta}))^2$ or the variance is proportional to a function of the mean response (here the square). This type of variance model is fairly typical when nonlinear data exhibit constant coefficient of variation (CV) rather than constant variance.

Therefore a variance function g is introduced that depends on $\boldsymbol{\beta}$ through the mean function f, on constants $\boldsymbol{z_j}$ that may include some or all of $\boldsymbol{x_j}$, and on possible extra variance parameters $\boldsymbol{\theta}$:

$$E(y_j) = f(x_j, \boldsymbol{\beta}); Var(y_j) = \sigma^2 g^2(\mu, \boldsymbol{z_j}, \boldsymbol{\theta}); \mu_j = f(\boldsymbol{x_j}, \boldsymbol{\beta}) \tag{2.7}$$

It can be difficult to specify values for $\boldsymbol{\theta}$ *a priori* and in such cases a data driven approach can be used. The choice of g may not require extra parameters i.e. when the variance is proportional to the square of the mean then θ is fixed at 2.

Intra-individual correlation The assumption that the e_j's are uncorrelated is replaced with a description of the assumed correlation pattern $\boldsymbol{\Gamma}(\boldsymbol{\alpha})$, where the correlation matrix $\boldsymbol{\Gamma}(\boldsymbol{\alpha})$ is a function of a vector of correlation parameters $\boldsymbol{\alpha}(s \times 1)$. Note that a scarcity of information may preclude reliable inference on within-individual correlation patterns.

Note that random effects can account for the correlation between measurements on a subject. While there might still be some pattern of correlation remaining, it can be very difficult to estimate correctly and adds more complexity to the model. Davidian and Giltinan (Davidian and Giltinan, 1998) warn that caution must be exercised when attempting to set up explicit models for within-subject correlation.

One example of a correlation structure typically used over time is that of AR(1):

$$\boldsymbol{\Gamma}(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n-1} \\ & 1 & \alpha & \dots & \alpha^{n-2} \\ & & 1 & \dots & \vdots \\ & & & \ddots & \alpha \\ & & & & 1 \end{pmatrix} \rightarrow \text{usually when measures are over time.}$$

If the variance is constant then $Cov(e) = \sigma^2 \boldsymbol{\Gamma}(\alpha)$.

General Covariance Structure The general covariance structure deals with the case where both correlation and heteroscedasticity are present.

Intra-individual variance heterogeneity One way to account for non-constant variance is to model the response variance in a similar way to how one models the mean response: eg $Var(y_j) = \sigma^2 (f(\boldsymbol{x_j}, \boldsymbol{\beta}))^2$ or the variance is proportional to a function of the mean response (here the square). This type of variance model is fairly typical when nonlinear data exhibit constant coefficient of variation (CV) rather than constant variance.

Therefore a variance function g is introduced that depends on $\boldsymbol{\beta}$ through the mean function f, on constants $\boldsymbol{z_j}$ that may include some or all of $\boldsymbol{x_j}$, and on possible extra variance parameters $\boldsymbol{\theta}$:

$$E(y_j) = f(x_j, \boldsymbol{\beta}); Var(y_j) = \sigma^2 g^2(\mu, \boldsymbol{z_j}, \boldsymbol{\theta}); \mu_j = f(\boldsymbol{x_j}, \boldsymbol{\beta}) \tag{2.7}$$

It can be difficult to specify values for $\boldsymbol{\theta}$ *a priori* and in such cases a data driven approach can be used. The choice of g may not require extra parameters i.e. when the variance is proportional to the square of the mean then θ is fixed at 2.

Intra-individual correlation The assumption that the e_j's are uncorrelated is replaced with a description of the assumed correlation pattern $\boldsymbol{\Gamma}(\boldsymbol{\alpha})$, where the correlation matrix $\boldsymbol{\Gamma}(\boldsymbol{\alpha})$ is a function of a vector of correlation parameters $\boldsymbol{\alpha}(s \times 1)$. Note that a scarcity of information may preclude reliable inference on within-individual correlation patterns.

Note that random effects can account for the correlation between measurements on a subject. While there might still be some pattern of correlation remaining, it can be very difficult to estimate correctly and adds more complexity to the model. Davidian and Giltinan (Davidian and Giltinan, 1998) warn that caution must be exercised when attempting to set up explicit models for within-subject correlation.

One example of a correlation structure typically used over time is that of AR(1):

$$\boldsymbol{\Gamma}(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n-1} \\ & 1 & \alpha & \dots & \alpha^{n-2} \\ & & 1 & \dots & \vdots \\ & & & \ddots & \alpha \\ & & & & 1 \end{pmatrix} \rightarrow \text{usually when measures are over time.}$$

If the variance is constant then $Cov(e) = \sigma^2 \boldsymbol{\Gamma}(\alpha)$.

General Covariance Structure The general covariance structure deals with the case where both correlation and heteroscedasticity are present.

Define the diagonal variance matrix $\boldsymbol{G}(\boldsymbol{\beta},\boldsymbol{\theta}) = diag[g^2(\mu_1, z_1, \boldsymbol{\theta}), \ldots, g^2(\mu_n, z_n, \boldsymbol{\theta})]$. Note that $\boldsymbol{\beta}$ is used here explicitly to emphasise the dependence on the regression parameters.

Then:

$$Cov(\boldsymbol{e}) = \sigma^2 \boldsymbol{G}^{\frac{1}{2}}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{\Gamma}(\boldsymbol{\alpha})\boldsymbol{G}^{\frac{1}{2}}(\boldsymbol{\beta},\boldsymbol{\theta}) \tag{2.8}$$

This can be written as $\boldsymbol{R}(\boldsymbol{\beta},\boldsymbol{\xi})$, where $\boldsymbol{\xi} = [\sigma, \boldsymbol{\theta}', \boldsymbol{\alpha}']$ which implies that: $Var(y_j) = \sigma^2 g^2(\mu_j, \boldsymbol{z_j}, \boldsymbol{\theta}), Corr(y_{j1}, y_{j2}) = \boldsymbol{\Gamma}_{j_1,j_2}(\boldsymbol{\alpha})$.

2.2.3 Hierarchical nonlinear models

This is an extension of the linear case where both intra-individual and inter-individual variation must be accounted for in a two stage model. In this case the first stage of intra-individual variation is characterised by a nonlinear function together with an individual covariance structure, and the second stage involves individual-specific regression parameters.

A fully parametric specification is presented here. One can get both semi-parametric and non-parametric specifications as well.

Stage 1: intra-individual For the first stage, f refers to a common function for all individuals but where the parameters of the function are allowed to vary for each individual:

$$\boldsymbol{y_i} = f_i(\boldsymbol{\beta_i}) + \boldsymbol{e_i} \tag{2.9}$$

where $\boldsymbol{f_i}(\boldsymbol{\beta_i}) = \begin{bmatrix} f(x_{i1}, \boldsymbol{\beta_i}) \\ \vdots \\ f(x_{in_i}, \boldsymbol{\beta_i}) \end{bmatrix}$

Paralleling the hierarchical linear model:
$E(\boldsymbol{e_i} \mid \boldsymbol{\beta_i}) = 0; Cov(\boldsymbol{e_i} \mid \boldsymbol{\beta_i}) = \boldsymbol{R_i}(\boldsymbol{\beta_i}, \boldsymbol{\xi})$ where $\boldsymbol{\xi} = [\sigma, \boldsymbol{\theta}', \boldsymbol{\alpha}']$ and hence

$$\boldsymbol{e_i} \mid \boldsymbol{\beta_i} \sim N(0, \boldsymbol{R_i}(\boldsymbol{\beta_i}, \boldsymbol{\xi})) \tag{2.10}$$

or a distribution other than the normal could be used. It is often convenient to write $\boldsymbol{R_i}(\boldsymbol{\beta_i}, \boldsymbol{\xi}) = \sigma^2 \boldsymbol{S_i}(\boldsymbol{\beta_i}, \gamma)$ where $\gamma = [\boldsymbol{\theta}', \boldsymbol{\alpha}']'$

This specification is very general and can accomodate both heterogeneous variance and within-subject correlations. The functional form for $\boldsymbol{R_i}$ is usually common to all subjects. Note that $\boldsymbol{R_i}$ is now more flexible then in the HLM

case and can now depend on covariates.

Stage 2: inter-individual The different values for the subject-specific regression parameters account for between subject or inter-individual variation. The second stage involves modelling these different parameters so that both systematic and random components are included.

The systematic components are captured through the vector of $\boldsymbol{\beta}$ that are common to all individuals and the random components are reflected with the random effects vector $\boldsymbol{b_i}$: $\boldsymbol{\beta_i} = \boldsymbol{\beta} + \boldsymbol{b_i}$; and with covariates included: $\boldsymbol{\beta_i} = \boldsymbol{A_i\beta} + \boldsymbol{b_i}$ where $\boldsymbol{A}$ is an indicator matrix in the case of group effects. Note that continuous covariates can also be included in this matrix.

The most general notation of this allows for random effects on certain parameters but not necessarily on others (through the design or indicator matrix $\boldsymbol{B}$): $\boldsymbol{\beta_i} = \boldsymbol{A_i\beta} + \boldsymbol{B_ib_i}$, or this can be written as:

$$\boldsymbol{\beta_i} = d(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}) \tag{2.11}$$

where $\boldsymbol{\alpha_i}$ is a $(\alpha \times 1)$ vector of covariate values for the ith subject and d is a vector-valued function with p dimensions, one for each regression parameter. The $\boldsymbol{b_i}$ are i.i.d. and often $b_i \sim (0, \boldsymbol{D})$.

Note that since the $\boldsymbol{\beta_i}$ in 2.11 are specific to particular individuals through the random effects $\boldsymbol{b_i}$:

$$\boldsymbol{e_i} \mid \boldsymbol{b_i} \sim (0, \boldsymbol{R_i}(\boldsymbol{\beta_i}, \boldsymbol{\xi})) \tag{2.12}$$

A new vector of parameters $\boldsymbol{\omega}$ then consists of both the intra-individual covariance parameters $\boldsymbol{\xi}$ and the distinct elements of the inter-individual covariance matrix $\boldsymbol{D}$.

This inter-individual stage can also be nonlinear, for example:
$\beta_{1i} = \beta_1 \exp(b_{1i})$ or with covariates: $\beta_{1i} = (\beta_1 + \beta_4 w_i)\exp(b_{1i})$ with w_i being weight. Other extensions to this setup include modelling time varying covariates, usually by allowing the β_i's to vary across time with the covariate of interest, and modelling multiple responses at once.

2.3 Inference and Estimation

It is important to incorporate fixed and random effects correctly in estimation and inference. Standard inferential techniques for nonlinear regression are based on the usual least squares principle as in the linear case. One can take 3 different views on the Ordinary Least Squares (OLS) procedure:

1. The OLS estimator $\hat{\beta}_{OLS}$ is equivalent to the maximum likelihood estimator under the classical assumptions. Under this view the usual likelihood principles are applied and the estimators have standard optimal properties.

2. A second view is that OLS is a more general method regardless of the distribution of the data. Under the assumption of independence, and with the general mean-variance specification, it appears reasonable to minimise the sum of squared deviations. If variance is constant then all deviations have the same importance whereas if variance is nonconstant then the deviations can be weighted in inverse proportion to the degree of uncertainty associated with the particular deviation.

3. One can also use OLS by solving the estimating equations according to some optimal criteria for β as opposed to minimising an objective function.

Therefore it is apparent from the second and third views of OLS that in order to apply OLS one only has to know the form of both the mean and variance (the first two moments). OLS estimation for nonlinear models is entirely analagous to the linear case but the fundamental difference is that the estimating equations cannot be solved explicitly and hence numerical methods must be used. When there is nonconstant variance then either Weighted Least Squares (WLS) or Generalised Least Squares (GLS) is usually used instead of OLS.

Least-squares estimators of linear models are unbiased, normally distributed, and have the minimum possible variance. The normality allows confidence intervals to be constructed. In comparison, nonlinear least squares estimators do not have these properties and only achieve them in asymptotia. The behaviour of these nonlinear least squares estimators varies greatly between different nonlinear models. Models where the nonlinear estimators behave similarly to linear estimators can be termed *close-to-linear* and conversely models that do not have this property can be called *far-from-linear.*

There are 3 different types of potential inference:

1. Inference on the fixed effects, for example a group effect

2. Inference on the variance components, for example the amount of within-subject variability relative to the between-subject variability

3. Inference on the subject means i.e. the systematic mean plus any individual random effect

2.3.1 Hierarchical linear models

This section begins by presenting the theory for a general linear mixed effects model before examining the interdependence of both fixed and random effects in the estimation procedure.

The general linear mixed effects model Combining the model for all m individuals leads to the following specification:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y_1} \\ \vdots \\ \boldsymbol{y_m} \\ (N \times 1) \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b_1} \\ \vdots \\ \boldsymbol{b_m} \\ (km \times 1) \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_1} \\ \vdots \\ \boldsymbol{X_m} \\ (N \times p) \end{bmatrix}$$

For m individuals there are a total of N or $\sum_{i=1}^{m} n_i$ responses, k is the # of random effects and p is the number of fixed effects.

Define block diagonal matrices $\tilde{\boldsymbol{D}} = diag(\boldsymbol{D}, \ldots, \boldsymbol{D})$, $\boldsymbol{Z} = diag(\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_m})$ and $\boldsymbol{R} = diag(\boldsymbol{R_1}, \ldots, \boldsymbol{R_m})$ so that $\boldsymbol{V} = diag(\boldsymbol{V_1}, \ldots, \boldsymbol{V_m}) = \boldsymbol{R} + \boldsymbol{Z}\tilde{\boldsymbol{D}}\boldsymbol{Z}'$.

The combined model can be written as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{e} \tag{2.13}$$

with the marginal distribution for the combined $\boldsymbol{y}$ vector $\sim N(\boldsymbol{X\beta}, \boldsymbol{V})$.

If the variance components ($\boldsymbol{V_i}$) are known then inference on $\boldsymbol{\beta}$ & $\boldsymbol{b_i}$ can be based on the marginal likelihood or the joint minimisation of the objective function:

$$log|\tilde{\boldsymbol{D}}| + \boldsymbol{b}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{b} + log|\boldsymbol{R}| + (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb}) \tag{2.14}$$

This may be accomplished using a 'pseudo-data' regression approach with the augmented response vector $\begin{bmatrix} \boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{y} \\ \boldsymbol{0}_{km \times 1} \end{bmatrix}$, where the dimensions of the $\boldsymbol{0}$ sub-vector correspond to the dimensions of the vector of random effects $\boldsymbol{b}$, and leads to the mixed-model equations:

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \tilde{\boldsymbol{D}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix} \tag{2.15}$$

Note that if one removed the $\tilde{\boldsymbol{D}}^{-1}$ from the left hand side then these equations would provide the maximum likelihood estimates with $\boldsymbol{b}$ being treated as

fixed effects.

This then solves to the following generalised least squares estimator:

$$\hat{\beta} = \left(\sum_{i=1}^{m} \boldsymbol{X_i}' \boldsymbol{V_i}^{-1} \boldsymbol{X_i} \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X_i}' \boldsymbol{V_i}^{-1} \boldsymbol{y_i} \tag{2.16}$$

and:

$$\hat{\boldsymbol{b}} = \tilde{\boldsymbol{D}} \boldsymbol{Z}' \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{2.17}$$

with the individual estimates

$$\hat{\boldsymbol{b}}_i = \boldsymbol{D} \boldsymbol{Z_i}' \boldsymbol{V_i}^{-1} (\boldsymbol{y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}) \tag{2.18}$$

As $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$ are linear functions of $\boldsymbol{y}$, standard errors can easily be calculated.

When the variance components are unknown, estimates of $\boldsymbol{\omega}$ are used and hence $\boldsymbol{V_i}$ would be replaced by $\hat{\boldsymbol{V}}_i$. Note that the standard errors are then a little low because they ignore the uncertaintly from estimates being used.

The limitations of this model is that both the random errors e_j and the b_i are assumed to be normal, as well as the assumption of linearity in the relationship between the response and the covariates.

Estimation of variance components In the linear case, the variance components contained in the vector $\boldsymbol{\omega}$ could be estimated using either Maximum Likelihood (ML) or Restricted Maximum Likelihood(REML).

Maximum likelihood involves maximising the the marginal loglikelihood with respect to the variance parameters $\boldsymbol{\omega}$:

$$LogL = -\frac{1}{2} N log 2\pi - \frac{1}{2} log|\boldsymbol{V}| - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X\beta})' \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) \tag{2.19}$$

Note that jointly maximising with respect to both $\boldsymbol{\beta}$ and the variance components leads to the GLS estimates of $\boldsymbol{\beta}$.

REML accounts for the loss in degrees of freedom due to the need to initially estimate $\boldsymbol{\beta}$. It is analagous to estimation using the residuals from the fit with fixed effects only and corresponds to maximising the following loglikelihood:

$$LogL_R = LogL + \frac{1}{2} p \; log 2\pi - \frac{1}{2} log|\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}| \tag{2.20}$$

where logL is evaluated at $\hat{\beta}$.

An iterative scheme would be used since we need estimates of the variance components in order to estimate the fixed effects, and we also need the fixed effects in order to estimate the variance components. One method of estimating the variance components is by the EM algorithm.

Examining this process for a simple mixed model ($y_{ij} = \mu + b_i + e_{ij}$) helps to illustrate some of the concepts related to estimating both fixed and random effects.

Estimation of fixed effects for known variance components It can be shown that the variance of the individual means $Var(\bar{y}_i) = \sigma_b^2 + \frac{\sigma_e^2}{n_i}$. Hence an estimate of the overall mean across the i subjects could be a weighted combination of the subject means, where the weights are inversely proportional to the variance of that subject mean:

$$\hat{\mu}_w = \frac{\sum_{i=1}^{m} w_i \bar{y}_i}{\sum_{i=1}^{m} w_i} \tag{2.21}$$

where $w_i = Var(\bar{y}_i)^{-1} = (\sigma_b^2 + \frac{\sigma_e^2}{n_i})^{-1}$

This is the GLS estimator of the overall mean μ and has minimum variance among all possible estimators that are based on weighted averages of the subject means.

But typically the variance components are unknown and so must be estimated. Hence estimation of fixed and random effects are inextricably linked together and so it is necessary to estimate the variance components in order to make valid inference on fixed effects i.e. even when one is not interested in the variance components themselves.

Estimation of random effects The random effects or the $\boldsymbol{b_i}$'s also need to be estimated in order to estimate the subject-specific means: $\mu_i = \mu + b_i$. A subject's mean $\bar{y}_{i.}$ provides information on the b_i whereby if $\bar{y}_{i.}$ is greater than the overall mean ($\bar{y}_{..}$) then it is likely that b_i is positive for that subject.

One can therefore predict a particular b_i by $\tilde{b}_i = E(b_i \mid \bar{y}_{i.})$. Note that the mean for the unconditional distribution of the b_i is zero and hence is not a reasonable predictor. It is straightforward to show:

$$\begin{bmatrix} b_i \\ \bar{y}_{i.} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix} \begin{bmatrix} \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \frac{\sigma_e^2}{n_i} \end{bmatrix} \right)$$

and from this the following expression can be derived:

$$E(b_i \mid \bar{y}_{i.}) = \frac{n_i \sigma_b^2}{\sigma_e^2 + \sigma_b^2 n_i}(\bar{y}_{i.} - \mu) \tag{2.22}$$

Now if μ is replaced by its GLS estimate, one gets the *best linear unbiased predictor* (BLUP) of b_i. Therefore the estimate of the mean for the ith subject is: $BLUP(\mu + b_i) = GLS(\mu) + BLUP(b_i)$ and this can be rewritten as: $BLUP(\mu + b_i) = \bar{y}_{i.} - \frac{\sigma_e^2}{n_i \sigma_b^2 + \sigma_e^2}(\bar{y}_{i.} - GLS(\mu))$

Examining this expression reveals that the subject mean $\bar{y}_{i.}$ is shrunk towards the grand mean μ and the extent of this shrinkage depends on the number of observations n_i for that subject and the relative magnitude of the two variance components. For example if n_i is large for subject i then the extent of the shrinkage will be small, or alternatively if σ_e^2 or within-subject variation is very large relative to σ_b^2 then the extent of the shrinkage will be large.

2.3.2 Nonlinear models for an individual

Generalised Least Squares (GLS) takes heteroscedasticity into account and is hence suitable for many nonlinear datasets. It can be used without knowledge of the distribution of the data, all one needs is to specify the first two moments i.e. the form of the mean and the variance.

When the variances of the y_j's are known up to a constant of proportionality i.e. $Var(y_j) = \frac{\sigma^2}{w_j}$ (for known w_j's) this leads to Weighted Least Squares (WLS). However usually one doesn't know this and one approach is to take advantage of the functional form for the variance to construct estimated weights: $\hat{w}_j = \frac{1}{g^2(\boldsymbol{x_j}, \hat{\boldsymbol{\beta}}_{OLS})}$

Under the general mean-variance specification (2.7) with $\boldsymbol{\theta}$ known, the GLS methods works as follows:

1. estimate $\hat{\boldsymbol{\beta}}$ using a preliminary estimator ($\hat{\boldsymbol{\beta}}_p$) such as from OLS
2. form estimated weights $\hat{w}_j = \frac{1}{g^2(\hat{\mu}_j, \boldsymbol{x_j}, \boldsymbol{\theta})}; \hat{\mu}_j = f(x_j, \hat{\boldsymbol{\beta}}_p)$
3. use these weights to re-estimate $\boldsymbol{\beta}$ by WLS. Return to step two and repeat.

The scale parameter σ^2 is estimated using $\hat{\boldsymbol{\beta}}_{GLS}$ i.e. from the final $\boldsymbol{\beta}$ estimates:

$$\hat{\sigma}^2_{GLS} = \frac{1}{n-p} \sum_{j=1}^{n} \hat{w}_j \left(y_j - f(\boldsymbol{x_j}; \hat{\boldsymbol{\beta}}_{GLS})\right)^2 \tag{2.23}$$

When one doesn't know $\boldsymbol{\theta}$ then it is usually estimated. One then replaces step two of the GLS algorithm by estimating $\hat{\boldsymbol{\theta}}$ and calculating weights.

GLS principles can be extended to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ in the general covariance specification for $\boldsymbol{R}$ where as before, $\boldsymbol{R}(\boldsymbol{\beta}, \boldsymbol{\xi})$ can be written as $\sigma^2 S(\boldsymbol{\beta}, \boldsymbol{\gamma})$; $\boldsymbol{\gamma} = \left[\boldsymbol{\theta}', \boldsymbol{\alpha}'\right]'$.

The algorithm proceeds as follows:

1. estimate $\boldsymbol{\beta}$ with a preliminary estimator $\hat{\boldsymbol{\beta}}_p$
2. estimate $\hat{\gamma}$ and form the estimated weight matrix based on $\hat{\boldsymbol{\beta}}_p$ and $\hat{\gamma}$: $\hat{\boldsymbol{W}} = S^{-1}(\hat{\boldsymbol{\beta}}_p, \hat{\gamma})$
3. Use $\hat{\boldsymbol{W}}$ to re-estimate $\boldsymbol{\beta}$. Go to step two and repeat...

The final estimate for σ^2 is based on the final estimates for $\boldsymbol{\beta}$ & $\boldsymbol{\gamma}$:

$$\hat{\sigma}^2_{GLS} = \frac{1}{n-p}(\boldsymbol{y} - f(\hat{\boldsymbol{\beta}}_{GLS}))' \boldsymbol{S}^{-1}(\hat{\boldsymbol{\beta}}_{GLS}, \hat{\boldsymbol{\gamma}})(\boldsymbol{y} - f(\hat{\boldsymbol{\beta}}_{GLS})) \tag{2.24}$$

Variance components estimation

There are several methods to estimate $\hat{\boldsymbol{\theta}}$. The methods presented below are all based on a transformation of absolute residuals from a preliminary fit and hence the objective functions depend on the data through these residuals $\boldsymbol{r_j}$. All these methods result in estimating equations that have a form similar to that of weighted least squares and hence only the first two moments are required:

- Psuedo Likelihood (PL) - PL minimises a function that corresponds to maximising the normal log likelihood evaluated at a preliminary estimate of $\hat{\boldsymbol{\beta}}_p$. This method doesn't take account of the loss of degrees of freedom due to the preliminary estimation of $\hat{\boldsymbol{\beta}}_p$.
- REML - this is unbiased as it accounts for the loss of degrees of freedom due to the preliminary estimation of $\hat{\boldsymbol{\beta}}_p$. Note that the development of this is analagous to the likelihood presented for the REML procedure in 2.20, except that the 1st term is now no longer the ML likelihood *LogL* but rather is the PL likelihood.

Note that variance components are considerably more difficult to estimate than regression parameters since information on higher moments are typically not as good as for lower moments.

Confidence Intervals & Hypothesis Testing In a linear model, under normality it is possible to derive exact distributional results for the OLS estimator for $\boldsymbol{\beta}$ and the unbiased estimator for σ^2. Even if the response is not normal the estimators are unbiased and the covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$ unchanged.

In a nonlinear model one cannot obtain exact, fixed sample size results due to the inability to solve the estimating equations explicitly. One can develop approximations using asymptotic theory and these asymptotic approximations hold even when normality is violated.

Under classic assumptions:
$\hat{\boldsymbol{\beta}}_{OLS} \sim N(\boldsymbol{\beta}, \sigma^2 \boldsymbol{\Sigma}_{OLS})$ where $\boldsymbol{\Sigma}_{OLS}{}^{-1} = \boldsymbol{X}'(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$

When variance is not constant:
$\hat{\boldsymbol{\beta}}_{GLS} \sim N(\boldsymbol{\beta}, \sigma^2 \boldsymbol{\Sigma}_{GLS})$ where $\boldsymbol{\Sigma}_{\boldsymbol{GLS}}{}^{-1} = \boldsymbol{X}'(\boldsymbol{\beta})\boldsymbol{S}^{-1}(\boldsymbol{\beta}, \boldsymbol{\gamma})\boldsymbol{X}(\boldsymbol{\beta})$

Under this notation, $\boldsymbol{X}(\boldsymbol{\beta})$ is the $(n \times p)$ matrix with the jth row equal to $\boldsymbol{f}'_{\beta}(\boldsymbol{x_j}, \boldsymbol{\beta})$. In the linear case $\boldsymbol{f}'(\boldsymbol{x_j}, \boldsymbol{\beta}) = \boldsymbol{x_j}'\boldsymbol{\beta}$ and the matrix $\boldsymbol{X}(\boldsymbol{\beta})$ is the usual design matrix.

'Sandwich' covariance estimators are robust to misspecifications of variance and are used in place of $\hat{\boldsymbol{\Sigma}}_{OLS}$ and $\hat{\boldsymbol{\Sigma}}_{GLS}$ to protect against improper modelling of heteroscedasticity. Note however that these estimators are highly sensitive to outliers.

Computational methods

Estimates for nonlinear least squares can rarely be found in closed form, iterative numerical methods are usually needed:

1. Newton-Raphson technique:
   If we are interested in maximising a scalar-valued objective function called $O(\boldsymbol{\tau})$ where $\boldsymbol{\tau}$ is a $(t \times 1)$ vector of parameters, an approximation may be derived using a quadratic Taylor expansion:

   $$O(\boldsymbol{\tau}) \approx O(\boldsymbol{\tau}^*) + \boldsymbol{s}'(\boldsymbol{\tau}^*)(\boldsymbol{\tau}^* - \boldsymbol{\tau}) + \tfrac{1}{2}(\boldsymbol{\tau}^* - \boldsymbol{\tau})'\boldsymbol{J}(\boldsymbol{\tau})(\boldsymbol{\tau}^* - \boldsymbol{\tau})$$

   where $\boldsymbol{s}(\boldsymbol{\tau})$ is the $(t \times 1)$ vector of partial derivatives of O with respect to the components of $\boldsymbol{\tau}$ (gradient vector), and $\boldsymbol{J}(\boldsymbol{\tau})$ is the $(t \times t)$ matrix of second derivatives (Hessian matrix).

   Maximising this expression leads to the following:

$\tau = \tau^* - J^{-1}(\tau^*)s(\tau^*)$ which lends itself to an iterative method of finding the values of τ that maximise O:

$$\tau^{(h+1)} = \tau^{(h)} - J^{-1}(\tau^{(h)})s(\tau^{(h)})$$

2. Fisher scoring (Gauss Newton) - The Hessian matrix can be difficult to compute and the Fisher Scoring method deals with this by replacing the $J(\tau)$ Hessian matrix by its expectation.

Either of these algorithms would continue to iterate until the difference in the objective function between successive iterations is small enough.

2.3.3 Hierarchical nonlinear models

There are 3 broad methods of inference for repeated measures nonlinear data. The first involves constructing individual-specific parameters and then uses these to get population averages, the second linearises the nonlinear function, and the 3rd method approximates the likelihood function using either a Laplacian approximation, or an adaptive Gaussian quadrature rule to refine this Laplacian approximation.

The first method requires sufficient richness in the data in order to obtain decent regression estimates for each subject yet often the data cannot support this. The second method involves approximating the nonlinear function with one that is additive in the random effects and random errors. The parameters are then estimated in a way analagous to the linear case. Since the software program used for this research (R 2.2, (R Development Core Team, 2005)) uses linearisation, this is the method that will be emphasised.

First-order linearization

In the hierarchical linear model scenario one can calculate the marginal distribution of the y_i's because both the random effects b_i and the individual errors e_i enter the model in an additive linear fashion, and have the assumption of independence from each other. The y_i will be normally distributed when the assumption of normality applies to both the b_i and the e_i.

The approximation below was suggested by Beal and Sheiner (1982) and uses the above information as follows:

Begin with the standard specification for a nonlinear model:

$$y_i = f_i(\beta_i) + e_i \; ; \; e_i \mid \beta_i \sim N(0, R_i(\beta_i, \xi)) \; ; \; \beta_i = d(\alpha_i, \beta, b_i).$$

Then let $\boldsymbol{e_i} = \boldsymbol{R_i}^{\frac{1}{2}}(\boldsymbol{\beta_i}, \boldsymbol{\xi})\boldsymbol{\epsilon_i}$ where:
$\boldsymbol{\epsilon_i} \sim (0, \boldsymbol{I_{n_i}})$ and are independent of the random effects $\boldsymbol{b_i}$'s; and $\boldsymbol{R_i}^{\frac{1}{2}}(\boldsymbol{\beta_i}, \boldsymbol{\xi})$ is the Cholesky Decomposition (square root of a matrix) of $\boldsymbol{R_i}$.

Stage 1 of the model can then be written as:
$\boldsymbol{y_i} = \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})\} + \boldsymbol{R_i}^{\frac{1}{2}}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}), \boldsymbol{\xi}\}\boldsymbol{\epsilon_i}$ and $\boldsymbol{\beta_i}$ is now written as a function $\boldsymbol{d}$ to explicitly emphasise the dependence on the random effects $\boldsymbol{b_i}$.

Then a Taylor Series expansion of the above expression is taken about the mean value $E(\boldsymbol{b_i}) = \boldsymbol{0}$, and keeping the first two terms in the expansion of $\boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})\}$ and the first term in the expansion of $\boldsymbol{R_i}^{\frac{1}{2}}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}), \boldsymbol{\xi}\}\boldsymbol{\epsilon_i}$ gives:

$$\boldsymbol{y_i} \approx \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{0})\} + \boldsymbol{F_i}(\boldsymbol{\beta}, \boldsymbol{0})\boldsymbol{\Delta_{bi}}(\boldsymbol{\beta}, \boldsymbol{0})\boldsymbol{b_i} + \boldsymbol{R_i}^{\frac{1}{2}}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{0}), \boldsymbol{\xi}\}\boldsymbol{\epsilon_i} \quad (2.25)$$

where $\boldsymbol{F_i}(\boldsymbol{\beta}, \boldsymbol{0})$ is the $(n_i \times p)$ matrix of derivatives of $\boldsymbol{f_i}(\boldsymbol{\beta_i})$ with respect to $\boldsymbol{\beta_i}$ and evaluated at $\boldsymbol{\beta_i} = \boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{0})$, and $\boldsymbol{\Delta_{bi}}(\boldsymbol{\beta}, \boldsymbol{0})$ is the $(p \times k)$ matrix of derivatives of $\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{0})$ with respect to $\boldsymbol{b_i}$ and evaluated at $\boldsymbol{b_i} = 0$.

Note that the approximation only takes the first term for the within-subject error as opposed to the first two terms for the mean function. This is justified due to the fact that misspecification of first moment properties is deemed more serious compared to second moment properties.

Examining (2.25) one can see that this is in a similar form to that of the linear case if one considers the last term as $\boldsymbol{e_i}^*$ and the $\boldsymbol{F_i}(\boldsymbol{\beta}, \boldsymbol{0})\boldsymbol{\Delta_{bi}}(\boldsymbol{\beta}, \boldsymbol{0})$ term as $\boldsymbol{Z_i}(\boldsymbol{\beta}, \boldsymbol{0})$:

$$\boldsymbol{y_i} \approx \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{0})\} + \boldsymbol{Z_i}(\boldsymbol{\beta}, \boldsymbol{0})\boldsymbol{b_i} + \boldsymbol{e_i}^* \quad (2.26)$$

So now both the $\boldsymbol{b_i}$ and the $\boldsymbol{e_i}^*$ enter the model in an additive manner. Note that there are a few important differences: $\boldsymbol{Z}$ is not a fixed design matrix here but depends on $\boldsymbol{\beta}$, the first term or fixed part of the model is nonlinear in $\boldsymbol{\beta}$, and lastly the covariance matrix of $\boldsymbol{e_i}^*$ is now a function of covariates $\boldsymbol{\alpha_i}$, fixed effects $\boldsymbol{\beta}$, and variance parameters $\boldsymbol{\xi}$.

The outcome of all this is that the marginal mean and covariance of $\boldsymbol{y_i}$ can be specified as:

$$E(\boldsymbol{y_i}) \approx \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{0})\},$$

$$Cov(\boldsymbol{y_i}) \approx \boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{0}), \boldsymbol{\xi}\} + \boldsymbol{Z_i}(\boldsymbol{\beta}, \boldsymbol{0})\boldsymbol{D}\boldsymbol{Z_i}'(\boldsymbol{\beta}, \boldsymbol{0}) \equiv \boldsymbol{V_i}(\boldsymbol{\beta}, \boldsymbol{0}, \boldsymbol{\omega})$$

Note the difference in $Cov(\boldsymbol{y_i})$ from the linear case i.e. now it also depends on the fixed effects.

If one now assumes that the approximation along with the first two moments are exact, either maximum likelihood or generalised least squares can be used to estimate the fixed effects $\boldsymbol{\beta}$ and the variance parameters $\boldsymbol{\omega}$.

Maximum Likelihood Under the assumption that $\boldsymbol{b_i}$ and $\boldsymbol{e_i^*}$ are normally distributed, one can obtain joint maximum likelihood estimates of both $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ by minimising twice the negative marginal normal likelihood of (2.26) :

$$\boldsymbol{L_{FO}(\beta,\omega)} = \sum_{i=1}^{m} \left(log|\boldsymbol{V_i(\beta,0,\omega)}| + [\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i},\boldsymbol{\beta},\boldsymbol{0})\}]' \boldsymbol{V_i}^{-1}(\boldsymbol{\beta},\boldsymbol{0},\boldsymbol{\omega})[\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i},\boldsymbol{\beta},\boldsymbol{0})\}] \right) \tag{2.27}$$

One of the numerical techniques such as the Newton-Raphson algorithm would be used in this minimisation. Often the OLS estimates (also called the 'naive' estimates) are used as starting values for the iterations.

Inference would be based on standard asymptotic theory i.e. standard errors would be obtained from the inverse of the information matrix evaluated at the estimates. Generally likelihood ratio tests are used to assist in choosing between nested models, and the Akaike information criterion (AIC) is used when comparing non-nested models.

Generalised least squares (GLS) The primary appeal of GLS is the fact that ML is sensitive to both potential non-normality in the response, and to misspecification of the individual covariance structure. Due to the complexity of the error term in (2.26) there are various ways to implement the GLS approach.

The basic way is:

- estimate $\boldsymbol{\beta}$ using a preliminary estimator $(\hat{\boldsymbol{\beta}}_p)$ such as from OLS
- estimate the variance parameters $\boldsymbol{\omega}$ by $\hat{\boldsymbol{\omega}}$ and form the estimated weight matrices: $\boldsymbol{W_i}(\hat{\boldsymbol{\beta}}_p, \boldsymbol{0}, \hat{\boldsymbol{\omega}})$
- use these weights to re-estimate $\boldsymbol{\beta}$ by WLS i.e. either minimisation of:

  $\Sigma_{i=1}^{m} [\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i},\boldsymbol{\beta},\boldsymbol{0})\}]' \boldsymbol{W_i}(\hat{\boldsymbol{\beta}}_p, \boldsymbol{0}, \hat{\boldsymbol{\omega}})[\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i},\boldsymbol{\beta},\boldsymbol{0})\}]$

  or equivalently solve the set of estimating equations:

$$\sum_{i=1}^{m} \boldsymbol{X_i}'(\boldsymbol{\beta}, \mathbf{0}) \boldsymbol{W_i}(\hat{\boldsymbol{\beta}}_p, \mathbf{0}, \hat{\boldsymbol{\omega}})[\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \mathbf{0})\}] = \mathbf{0}$$

where $\boldsymbol{X_i}(\boldsymbol{\beta}, \mathbf{0}) = \boldsymbol{F_i}(\boldsymbol{\beta}, \mathbf{0}) \boldsymbol{\Delta_{\beta i}}(\boldsymbol{\beta}, \mathbf{0})$, and $\boldsymbol{\Delta_{\beta i}}(\boldsymbol{\beta}, \mathbf{0})$ is as before, i.e. a matrix of derivatives of $\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})$ with respect to $\boldsymbol{\beta}$ evaluated at $\boldsymbol{b_i} = \mathbf{0}$.

- Return to step two and repeat using the updated $\boldsymbol{\beta}$ estimates.

One would estimate $\boldsymbol{\omega}$ in step two with either PL or REML minimising the respective objective functions:

- PL: $\sum_{i=1}^{m} PL_i(\hat{\boldsymbol{\beta}}_p, \boldsymbol{\omega})$ where
  $PL_i(\hat{\boldsymbol{\beta}}_p, \boldsymbol{\omega}) = [\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \hat{\boldsymbol{\beta}}_p, \mathbf{0})\}]' \boldsymbol{V_i}^{-1}(\hat{\boldsymbol{\beta}}_p, \mathbf{0}, \boldsymbol{\omega}) \times [\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \hat{\boldsymbol{\beta}}_p, \mathbf{0})\}] + log|\boldsymbol{V_i}(\hat{\boldsymbol{\beta}}_p, \mathbf{0}, \boldsymbol{\omega})|$
- REML: $\sum_{i=1}^{m} REML_i(\hat{\boldsymbol{\beta}}_p, \boldsymbol{\omega})$ where
  $REML_i(\hat{\boldsymbol{\beta}}_p, \boldsymbol{\omega}) = PL_i(\hat{\boldsymbol{\beta}}_p, \boldsymbol{\omega}) + log|\boldsymbol{X_i}'(\hat{\boldsymbol{\beta}}_p, \mathbf{0}) \boldsymbol{V_i}^{-1}(\hat{\boldsymbol{\beta}}_p, \mathbf{0}, \boldsymbol{\omega}) \boldsymbol{X_i}'(\hat{\boldsymbol{\beta}}_p, \mathbf{0})|$

Just like in the case of 1 individual, PL or REML estimation does not require normality but only assumptions about the first two moments.

Conditional first-order linearization

An important point about the first order linearisation presented above is that variation between individuals is not incorporated in the mean function and only enters the model through the additive linear term $\boldsymbol{Z_i}(\boldsymbol{\beta}, \mathbf{0}) \boldsymbol{b_i}$. If between individual variation is a large component of variability then the model (2.26) may be a poor approximation.

Davidian and Giltinan (1998) presents a procedure advocated by Lindstrom and Bates in 1990 that involves refining the linearisation so that the Taylor series expansion is no longer taken about the expectation of $\boldsymbol{b_i}$ being zero, but rather around some value $\boldsymbol{b_i}^*$ that is closer to $\boldsymbol{b_i}$ then its expectation. This leads to the following:

$$\boldsymbol{y_i} \approx \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}^*)\} + \boldsymbol{F_i}(\boldsymbol{\beta}, \boldsymbol{b_i}^*) \boldsymbol{\Delta_{bi}}(\boldsymbol{\beta}, \boldsymbol{b_i}^*)(\boldsymbol{b_i} - \boldsymbol{b_i}^*) + \boldsymbol{R_i}^{\frac{1}{2}}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}^*), \boldsymbol{\xi}\}\epsilon_i \tag{2.28}$$

$\boldsymbol{F_i}$ and $\boldsymbol{\Delta_{bi}}$ are defined as before except now the derivatives are evaluated at $\boldsymbol{b_i}^*$ rather than its expectation of $\mathbf{0}$.

Analagous to (2.26) this can then be written as:

$$\boldsymbol{y_i} \approx \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}^*)\} + \boldsymbol{Z_i}(\boldsymbol{\beta}, \boldsymbol{b_i}^*)\boldsymbol{b_i} - \boldsymbol{Z_i}(\boldsymbol{\beta}, \boldsymbol{b_i}^*)\boldsymbol{b_i}^* + \boldsymbol{e_i}^* \tag{2.29}$$

Therefore one can see that there is now an extra additive term: $-\boldsymbol{Z}(\boldsymbol{\beta}, \boldsymbol{b_i}^*)\boldsymbol{b_i}^*$.

The approximate moments of $\boldsymbol{y_i}$ are then:

$$E(\boldsymbol{y_i}) \approx \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}^*)\} - \boldsymbol{Z}(\boldsymbol{\beta}, \boldsymbol{b_i}^*)\boldsymbol{b_i}^*,$$

$$Cov(\boldsymbol{y_i}) \approx \boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}^*), \boldsymbol{\xi}\} + \boldsymbol{Z_i}(\boldsymbol{\beta}, \boldsymbol{b_i}^*)\boldsymbol{D}\boldsymbol{Z_i}'(\boldsymbol{\beta}, \boldsymbol{b_i}^*) \equiv \boldsymbol{V_i}(\boldsymbol{\beta}, \boldsymbol{b_i}^*, \boldsymbol{\omega})$$

and if $\boldsymbol{b_i}$ and $\boldsymbol{e_i}^*$ are normally distributed then the (approximate) marginal distribution will also be normal.

One needs a reasonable choice for $\boldsymbol{b_i}^*$ in order to find decent estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$.

GLS for conditional first order linearisation Davidian and Giltinan (1998) present an iterative GLS method that was proposed by Lindstrom and Bates for a restricted version of the hierarchical nonlinear model proposed where the inter-individual function $\boldsymbol{d}$ is linear in $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$, and where the intra-individual matrix $\boldsymbol{R_i}(\boldsymbol{\beta_i}, \boldsymbol{\xi})$ does not depend on $\boldsymbol{\beta_i}$ and so can be written as $\boldsymbol{R_i}(\boldsymbol{\xi})$. This method is generalised after the restricted version is presented below:

Assume that $\boldsymbol{\omega}(\boldsymbol{D}\ \&\ \boldsymbol{\xi})$ is known. Then one can obtain estimates for $\boldsymbol{\beta}$ & $\boldsymbol{b_i}$ based on the joint minimisation of the objective function:

$$\sum_{i=1}^{m} \left(log|\boldsymbol{D}| + \boldsymbol{b_i}'\boldsymbol{D}^{-1}\boldsymbol{b_i} + log|\boldsymbol{R_i}(\boldsymbol{\xi})| + [\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})\}]'\boldsymbol{R_i}^{-1}(\boldsymbol{\xi})[\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})\}]\right) \tag{2.30}$$

which is twice the negative log of the posterior density of $\boldsymbol{b_i}$ for fixed $\boldsymbol{\beta}$ and twice the negative log of the posterior density of $\boldsymbol{\beta}$ for fixed $\boldsymbol{b_i}$.

Therefore the method involves the following two stage approach whereby the first stage is a penalised least squares (PNLS) step, and the second is a linear mixed effects (LME) stage:

1. Estimate $\boldsymbol{\omega}$ with $\hat{\boldsymbol{\omega}}$ and then minimise the following function to obtain estimates for the random effects and the initial fixed effects denoted as $\hat{\boldsymbol{b}}_i$ and $\hat{\boldsymbol{\beta}}_0$ respectively:

   $$\textstyle\sum_{i=1}^{m} \left(log|\hat{\boldsymbol{D}}| + \boldsymbol{b_i}'\hat{\boldsymbol{D}}^{-1}\boldsymbol{b_i} + log|\boldsymbol{R_i}(\hat{\boldsymbol{\xi}})| + [\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})\}]\boldsymbol{R_i}^{-1}(\hat{\boldsymbol{\xi}})[\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})\}]\right)$$

2. Estimate $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ with $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\omega}}$ after minimising:

   $$L_{LB}(\boldsymbol{\beta}, \boldsymbol{\omega}) = \textstyle\sum_{i=1}^{m} \left(log|\boldsymbol{V_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i, \boldsymbol{\omega})| + \boldsymbol{r_i}^{*\prime}(\boldsymbol{\beta}, \hat{\boldsymbol{b}}_i, \hat{\boldsymbol{\beta}}_0)\boldsymbol{V_i}^{-1}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i, \boldsymbol{\omega})\boldsymbol{r_i}^*(\boldsymbol{\beta}, \hat{\boldsymbol{b}}_i, \hat{\boldsymbol{\beta}}_0)\right)$$
   where $\boldsymbol{r_i}^*(\boldsymbol{\beta}, \hat{\boldsymbol{b}}_i, \hat{\boldsymbol{\beta}}_0) = \boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \hat{\boldsymbol{b}}_i)\} + \boldsymbol{Z_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\hat{\boldsymbol{b}}_i$ i.e. includes a component for inter-individual variation in the mean response function.

   Alternatively one could take a REML approach and minimise:
   $L_{LB,REML} = L_{LB}(\boldsymbol{\beta}, \boldsymbol{\omega}) + log|\boldsymbol{X_i}'(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\boldsymbol{V_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i, \boldsymbol{\omega})\boldsymbol{X_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)|$
   where $\boldsymbol{X_i}(\boldsymbol{\beta}, \hat{\boldsymbol{b}}_i) = \boldsymbol{F_i}(\boldsymbol{\beta}, \hat{\boldsymbol{b}}_i)\boldsymbol{\Delta_{\beta i}}(\boldsymbol{\beta}, \hat{\boldsymbol{b}}_i)$

This algorithm would iterate until convergence and therefore would produce: $\hat{\boldsymbol{\beta}}_{LB}, \hat{\boldsymbol{\omega}}_{LB}$ & $\hat{\boldsymbol{b}}_{i,LB}$. Note that the REML likelihood depends on both the fixed and random effects and hence the REML likelihood for models with either different fixed effects or different random effect structures cannot be compared i.e. with a likelihood ratio test.

Step 1 is called a 'psuedo-data' step since the joint minimisation may be achieved simultaneously by setting up an augmented nonlinear least squares problem as follows:

As in section 2.3.1 define the matrices $\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{R}$, and $\tilde{\boldsymbol{D}}$ and let $\boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}) = [\boldsymbol{f}_1'\{\boldsymbol{d}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}, \boldsymbol{b}_1)\}, \ldots, \boldsymbol{f}_m'\{\boldsymbol{d}(\boldsymbol{\alpha}_m, \boldsymbol{\beta}, \boldsymbol{b}_m)\}]'$, $\hat{\boldsymbol{R}}_i = \boldsymbol{R}_i(\hat{\boldsymbol{\xi}})$ and $\hat{\boldsymbol{R}} = diag(\hat{\boldsymbol{R}}_1, \ldots, \hat{\boldsymbol{R}}_m)$ and $\hat{\tilde{\boldsymbol{D}}} = diag(\hat{\boldsymbol{D}}, \ldots, \hat{\boldsymbol{D}})$.

$$\text{Regress} \begin{bmatrix} \hat{\boldsymbol{R}}^{-\frac{1}{2}}\boldsymbol{y} \\ \boldsymbol{0}_{km\times 1} \end{bmatrix} \text{ on } \begin{bmatrix} \hat{\boldsymbol{R}}^{-\frac{1}{2}}\boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}) \\ \hat{\tilde{\boldsymbol{D}}}^{-\frac{1}{2}}\boldsymbol{b} \end{bmatrix}$$

This is equivalent to minimising a penalised nonlinear least squares objective function with the extra penalty term $||\tilde{\boldsymbol{D}}\boldsymbol{b}_i||^2$:

$\sum_{i=1}^{m} \left[||\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)||^2 + ||\tilde{\boldsymbol{D}}\boldsymbol{b}_i||^2 \right]$ but where the responses are weighted by $\hat{\boldsymbol{R}}^{-\frac{1}{2}}$. This step is the nonlinear version of the standard approach to estimating $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ i.e. nonlinear form of the linear mixed-model equations (2.15).

Step two uses the approximate moments of $\boldsymbol{y}_i$ (under conditional first-order linearization) to approximate the marginal distribution of $\boldsymbol{y}_i$ and solves this with either ML or REML.

$$E(\boldsymbol{y}_i) \approx \boldsymbol{f}_i\{\boldsymbol{d}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}, \hat{\boldsymbol{b}}_i)\} - \boldsymbol{Z}_i(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\hat{\boldsymbol{b}}_i,$$

$$Cov(\boldsymbol{y}_i) \approx \boldsymbol{R}_i(\boldsymbol{\xi}) + \boldsymbol{Z}_i(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\boldsymbol{D}\boldsymbol{Z}_i'(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i) \equiv \boldsymbol{V}_i(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i, \boldsymbol{\omega})$$

Note that the notation now has $\hat{\boldsymbol{b}}_i$ in place of $\boldsymbol{b}_i^*$ and $\hat{\boldsymbol{\beta}}_0$ in place of $\boldsymbol{\beta}$. Both of these are estimated in the first step. The matrix $\boldsymbol{Z}(\boldsymbol{\beta}, \boldsymbol{b}_i^*)$ is evaluated at these estimates and then is regarded as fixed thereby corresponding to the linear case where Z is a fixed design matrix. The resulting estimators can be viewed as GLS or psuedo-likelihood estimators.

This step is also called the 'linear mixed effects' step since one can use an additional approximation that allows the minimisation required in step two to be expressed in a linear form. Approximate $\boldsymbol{f}_i\{\boldsymbol{d}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}, \hat{\boldsymbol{b}}_i)\}$ with a linear function with respect to $\boldsymbol{\beta}$ by expanding this expression around the initial estimate $\hat{\boldsymbol{\beta}}_0$ from step 1:

$$\begin{aligned} \boldsymbol{r_i}^*(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i) &= \boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \hat{\boldsymbol{b}}_i)\} + \boldsymbol{Z_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\hat{\boldsymbol{b}}_i \\ &\approx \boldsymbol{y_i} - \left[\boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\} + \boldsymbol{X_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)\right] + \boldsymbol{Z_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\hat{\boldsymbol{b}}_i \\ &= \boldsymbol{y_i}^* - \boldsymbol{X_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\boldsymbol{\beta} \end{aligned}$$

where $\boldsymbol{y_i}^* = \boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\} + \boldsymbol{X_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i) + \boldsymbol{Z_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\hat{\boldsymbol{b}}_i$

Then if $\boldsymbol{y_i}^* - \boldsymbol{X_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\boldsymbol{\beta}$ is substituted for $\boldsymbol{r_i}^*(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)$ in step two of the algorithm, one ends up with a function similar in form to the linear case and computational methods for the hierarchical linear model can be used.

Now generalising this to the case where the intra-individual variance matrix depends on the fixed effects i.e. $\boldsymbol{R_i}(\boldsymbol{\beta_i}, \boldsymbol{\xi}) = \boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})$, similar to (2.30) the objective function with $\boldsymbol{\omega}$ known becomes:

$$\sum_{i=1}^{m}(log|\boldsymbol{D}| + \boldsymbol{b_i}'\boldsymbol{D}^{-1}\boldsymbol{b_i} + log|\boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}), \boldsymbol{\xi}\}| + [\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})\}]' \times \boldsymbol{R_i}^{-1}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}), \boldsymbol{\xi}\}[\boldsymbol{y_i} - \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i})\}]) \quad (2.31)$$

But now $\boldsymbol{R_i}(\boldsymbol{\beta_i}, \boldsymbol{\xi})$ can exhibit a nonlinear dependence on these parameters and hence minimising (2.31) with the 'psuedo-data' approach described above is not straight forward. Also note that since the intra-individual variance now depends on the fixed effects, the resulting estimator for $\boldsymbol{\beta}$ will not be equal to the GLS estimator.

One possible strategy is to begin by taking starting estimates for $\hat{\boldsymbol{b}}_i$ and $\hat{\boldsymbol{\beta}}_0$ and the marginal distribution of $\boldsymbol{y_i}$ is then assumed to be Normal with:

$$E(\boldsymbol{y_i}) \approx \boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \hat{\boldsymbol{b}}_i)\} - \boldsymbol{Z_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\hat{\boldsymbol{b}}_i,$$

$$Cov(\boldsymbol{y_i}) \approx \boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i), \boldsymbol{\xi}\} + \boldsymbol{Z_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i)\boldsymbol{D}\boldsymbol{Z_i}'(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i) \equiv \boldsymbol{V_i}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}}_i, \boldsymbol{\omega})$$

Note that $\boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}), \boldsymbol{\xi}\}$ is evaluated at the estimated values for $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{b}}_i$ and does not depend on the unknown $\boldsymbol{\beta}$ and hence is not really different from the treatment for $\boldsymbol{R}(\boldsymbol{\xi})$.

Therefore if $\boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \boldsymbol{\beta}, \boldsymbol{b_i}), \boldsymbol{\xi}\}$ is approximated by $\boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i}, \hat{\boldsymbol{\beta}}_{00}, \hat{\boldsymbol{b}}_{i,0}), \boldsymbol{\xi}\}$ where $\hat{\boldsymbol{\beta}}_{00}$ & $\hat{\boldsymbol{b}}_{i,0}$ are previous estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$, the following two step procedure can be used. Note that while this is based on further approximation, it is expected to perform comparably to the simpler case:

1. Estimate $\boldsymbol{\omega}$ with $\hat{\boldsymbol{\omega}}$ and use previous estimates $\hat{\boldsymbol{\beta}}_{00}$ & $\hat{\boldsymbol{b}}_{i,0}$ to minimise the following function to obtain estimates for the random effects $\hat{\boldsymbol{b}}_i$ and the

initial fixed effects $\hat{\boldsymbol{\beta}}_0$:

$$\sum_{i=1}^{m}(log|\hat{\boldsymbol{D}}|+\boldsymbol{b_i}'\hat{\boldsymbol{D}}^{-1}\boldsymbol{b_i}+log|\boldsymbol{R_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i},\hat{\boldsymbol{\beta}}_{00},\hat{\boldsymbol{b}}_{i,0}),\hat{\boldsymbol{\xi}}\}|+[\boldsymbol{y_i}-\boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i},\boldsymbol{\beta},\boldsymbol{b_i})\}]'$$
$$\boldsymbol{R_i}^{-1}\{\boldsymbol{d}(\boldsymbol{\alpha_i},\hat{\boldsymbol{\beta}}_{00},\hat{\boldsymbol{b}}_{i,0}),\hat{\boldsymbol{\xi}}\}[\boldsymbol{y_i}-\boldsymbol{f_i}\{\boldsymbol{d}(\boldsymbol{\alpha_i},\boldsymbol{\beta},\boldsymbol{b_i})\}])$$

2. Estimate $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ with $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\omega}}$ after minimising one of the functions ($L_{LB}(\boldsymbol{\beta},\boldsymbol{\omega})$ or $L_{LB,REML}(\boldsymbol{\beta},\boldsymbol{\omega})$) where $\boldsymbol{V_i}(\hat{\boldsymbol{\beta}}_0,\hat{\boldsymbol{b}}_i,\boldsymbol{\omega})$ is defined as earlier. Update $\hat{\boldsymbol{\beta}}_{00}$ with $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_{i,0}$ with $\hat{\boldsymbol{b}}_i$

The above process is then iterated until convergence is achieved. Note that step 1 may be implemented using the 'psuedo-data' approach and the estimators will have the form of the GLS estimators.

Inference is based on treating the approximate moments evaluated at $\hat{\boldsymbol{b}}_{i,LB}$ as exact and then using asymptotic theory. As before, inference is therefore only approximate plus there is no accounting for the initial estimation of the random effects.

2.4 Software used

There are several software packages that are capable of nonlinear mixed modeling. The best known of these packages are $NONMEM^{TM}$, PROC NLMIXED from SAS^{TM}, $WinNonmix^{TM}$, PKBUGS (WinBUGS), and the 'nlme' (Pinheiro et al., 2006) package in R (R Development Core Team, 2005) that was available to me for this research.

The nlme package (Pinheiro et al., 2006) uses the two step algorithm proposed by Lindstron and Bates and described above. It is implemented for maximum likelihood and for restricted maximum likelihood. The first stage is the penalised least squares step (PNLS) and involves holding the variance components $\boldsymbol{\omega}$ fixed and estimating the random effects $\boldsymbol{b_i}$ and initial fixed effects $\boldsymbol{\beta}_0$. Since one cannot solve this explicitly, the Fisher scoring or Gauss-Newton computational method is used.

The second step or the LME step then updates the estimate of the variance components $\boldsymbol{\omega}$ and the fixed effects $\boldsymbol{\beta}$ based on a first-order Taylor series expansion of the model function around the estimates of $\boldsymbol{\beta}_0$ and the conditional modes of the random effects $\boldsymbol{b_i}$ that are obtained from the first step.

Step-halving is used at each Gauss-Newton iteration to ensure that the updated parameter estimates leads to a decrease in the objective function. This

works by checking if the objective function has improved with the updated estimates. If the objective function has improved then iteration continues while if it does not improve one calculates a new updated estimate by taking the original estimate and adding half the increment to the new value. The process is repeated with the increment becoming smaller and smaller until a decrease is observed in the objective function or some predetermined step size is reached.

Chapter 3

Analysis

3.1 The Data

3.1.1 SEACAT evaluation

The South East African Combination Anti-malarial Therapy (SEACAT) evaluation received seed funding from the UNDP / World Bank / WHO Special Program for Research and Training in Tropical Diseases, to evaluate comprehensively the effect of the wide scale implementation of artemisinin-based combination therapy within the normal context of use as first line treatment of uncomplicated malaria in the public sector in Southern Africa.

The SEACAT evaluation assesses the phased introduction of Artemisinin-based Combination Therapy (ACT) at provincial or district level in southern Mozambique, Swaziland and South Africa. The ACT consists of artesunate plus SP in all study sites except for KwaZulu Natal where the high level of resistance to SP necessitates an alternative combination therapy of artemether plus lumefantrine. In vivo studies will be conducted prior to the introduction of ACT to evaluate the efficacy of SP monotherapy. These results will then be compared with those following ACT.

3.1.2 Focus of this study

This study focuses on the gametocyte density measurements in open label clinical trials of subjects treated with SP as first-line therapy for uncomplicated malaria at the Naas/Mangweni clinics (Mpumalanga) and Namaacha and Bela Vista Clinics (Mozambique). The study design focused on measuring the response of asexual parasites to treatment. Gametocyte densities were measured at the same time resulting in gametocyte densities being measured on days 0, 3, 7, 14, 21, 28, and 42.

Gametocyte density was counted on thick malaria smears against 1000 leukocytes assuming 8000 leukocytes per microlitre (μL). This explains why the lowest detectable density is 8 μL as the smallest possible count would be 1 gametocyte per 1000 leukocytes which would then get multiplied up by 8 to get to units in terms of a microlitre.

It is pertinent that patients that fail treatment (in that they don't clear their asexual parasites) are withdrawn from the study to be given rescue treatment and effectively lost to follow up. Due to the time lag between asexual parasites and sexual gametocytes, the distribution of gametocyte densities for these patients are often truncated or in some cases the patients are withdrawn prior to the emergence of gametocytes. Unfortunately this means that there is informative censoring in the data when it comes to treatment failures.

3.1.3 Description of the covariates

The covariates included in this analysis are described below. Note that day 0 indicates the start of treatment and that variable names are shown in brackets:

- Site (*site*): Note that there were only three patients included from the Bela Vista site. These patients were therefore included with the Namaacha site as a general Mozambique site leaving two sites in total. This was coded so that the reference level was for Mpumulanga.
- Age (*age*): Age was measured in years and is accepted as a proxy for immunity.
- Gender (*gender*): Males were compared to females.
- Logged day 0 parasite density (logpdens): Logged values (base10) of the measured asexual parasite density at baseline.
- Mutations (*mutcat2*): The mutations variable was *a priori* expected to play an important part as it indicates the degree of resistance that the parasite has to SP treatment. It was coded to contrast all five mutations versus the rest (i.e. fewer than five mutations). Note that the mutation variable *mutcat2* is actually labelled in the model building plots as *mut_3* but refers to the same variable.
- Treatment outcome (*fail*): This is the treatment outcome variable and compares patients who failed treatment (resistant) to those with an adequate clinical and parasitological response (sensitive).
- Parasite clearance time (*pct*): Asexual parasite clearance time measured in days from the start of treatment. There were only four unique values for this variable namely 1, 2, 3 and 7 days, as these were the days stipulated for routine follow up and failure to clear parasites by day 7 would have resulted in the patient being withdrawn from the study and given rescue treatment.

- Patient category (*pcat*): This refers to the empirical patient categorisation explained in section 3.1. Since these groupings cannot be explained biologically on the basis of the data at hand, this was the last covariate to enter the model regardless of the patterns seen in the plots. The rationale being to consider the biological information first. Note that this variable was modelled by setting up dummy variables comparing categories two and three with category one as the reference level.

There was also sporadic data for certain pharmacokinetic parameters relating to SP drug levels but these were ultimately ignored as incorporating them would have resulted in too much data loss.

3.2 Data Preparation

The data was received in several different files, some of which contained demographic, clinical and parasitological readings and others contained data on mutations. Stata 8.2 (StataCorp, College Station, Texas, 2005) was used to merge the data together into a single coherent dataset and the code for this is found in Appendix D in section D.1.1.

The original gametocyte readings were logged to the base two for the modeling. It is common to log a measurement that is characterised by long right tails and where it is more meaningful to express increases in a multiplicative rather than additive manner. This transformation was also necessary to stabilise the models. The base of two was chosen due to the fact that certain important parameters are interpreted as a doubling or halving of the response. For example the gametocyte elimination half-life is the time it takes for the density of gametocytes to halve. Using a logarithmic scale with base two allows a one unit change in the response to be viewed as a doubling or halving. Note that this was implemented as logging the original value plus one in order to accommodate the malaria smears on which no gamteocytes were present. Asexual parasite density at day 0 was logged to the base 10 i.e. a unit change represents a 10 fold increase/decrease.

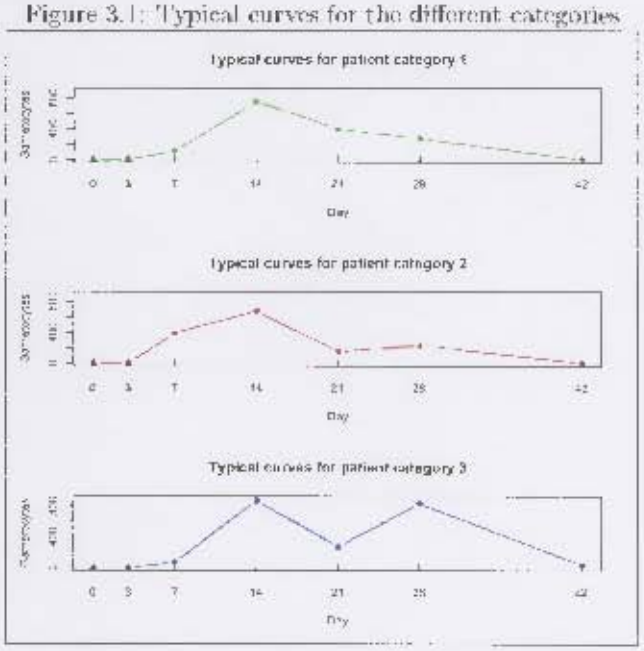
The main problem with the data was the high prevalence of zero measurements and their associated uncertainty. A zero reading does not definitely indicate no gametocytes, due to the fact that there is a detectable limit for the technology used i.e. the procedure cannot detect gametocytes if they are below a density of 8 gametocytes per microlitre (μL). Note that there is also error associated with these measurements due to the fact that only a portion of a bloodsmear is examined microscopically yet the gametocytes are not necessarily uniformly distributed across the bloodsmear.

Attempting to include a lot of zeros in the data when trying to solve a nonlinear mixed model leads to estimation difficulties. Therefore there were two

broad treatments of these zeros. In both cases the starting point or zero reading at day 0 was left as a zero and patients exhibiting gametocytes on day 0 were excluded. In the first dataset (hereafter referred to as *Data1*) all other zeros were simply made missing, while in the 2nd instance (*Data2*) they were changed to the detectable limit of 8 gametocytes per microlitre (μL).

Another important feature of the data was the fact that, broadly speaking, three different patterns in the untransformed response were observed across the sample of patients (see figure 3.1). A decision was made to split the datasets up according to these patterns thereby allowing one to assess the influence on estimates from excluding/including various patient groups.

Figure 3.1: Typical curves for the different categories



At this stage there was no proven biological basis for the grouping. Although it is possible that the appearance of a 2nd wave is related to synchronicity, none of the covariates that were measured in this study can be used to predict synchronicity e.g. immunity and duration of infection. The grouping of patients was therefore an empirical exercise and the categories were created as follows:

1. this was the largest patient category and included patients that experienced an initial increase followed by a decrease that did not increase again.

2. the 2nd group was for patients who experienced a second increase in gametocytes after the initial rise had started going down but the increase was slight i.e. less than a twofold increase.

3. the 3rd category included those patients that experienced a marked second increase in gametocytes after the initial rise had started going down. The marked increase was defined as a change of at least twofold i.e if a patient had exhibited decreasing gametocytes and then a measurement suddenly resulted in the previous value doubling (say from 75 to at least 150).

3.3 Data Exploration

Stata 8.2 (StataCorp, College Station, Texas, 2005) was used for the data exploration described below. The demographics of the sample, including patients that were excluded, are presented and then the high prevalence of zeros in the data is briefly discussed. The results of the data exploration to investigate possible correlates with the empirically created patient category variable, and also to investigate other relationships amongst the covariates are shown. The reason for doing this is that in addition to modeling gametocyte densities over time, the effect of demographic and disease-specific covariates on these density-time profiles will be examined. The corresponding Stata output can be found in Appendix D in section D.1.2.

3.3.1 Sample characteristics

There was a total of 579 patients recruited in SEACAT in vivo therapeutic efficacy studies for SP. Gametocytes were not detected on any day for 305 patients. Only patients who carried gametocytes were considered for this analysis as the objective was to model positive or nonzero gametocyte densities over time. Patients with gametocytes present at day 0 were excluded from the analysis as these gametocytes could not be cleared by SP treatment (since SP cannot effect mature gametocytes), and patients with insufficient gametocyte data were also excluded. In total there were 103 patients with sufficient gametocyte data that were used in the modeling.

Within each of the two datasets (formed according to the way zeros were handled) there were two different ways of selecting patients according to the minimum number of nonzero gametocyte readings (at least 3 and at least 4 readings). Combining this with the three patient categories resulted in 6 different *data sizes* in total. The different data sizes are displayed below in table 3.1. Note that the actual number of patients in each category is given in brackets while the numbers outside of the brackets are cumulative.

Table 3.1: Data sizes

Patient Group	Min no. +ve Readings	
	4	3
1	34 (34)	74 (74)
1+2	47 (13)	88 (14)
1+2+3	56 (9)	103 (15)

Table 3.2 summarises the demographic information for the full sample. It is presented in three 'sample groups': group one consists of the patients without any gametocytes, group two contains those patients that were excluded for one of the reasons given below, and group three consists of the patients that were included in the analysis.

Group two contained 23 patients that had gametocytes present at day 0 before treatment, 139 patients where gametocytes were detected on fewer than 3 observations, and nine patients who were from the Mpumulanga 1998 cohort that did not have actual gametocyte densities but rather absence/presence data.

Medians are reported for age and parasite clearance time as the distributions for these variables are positively skewed. Parasite density is typically reported as a geometric mean and hence that statistic has been used in the table. Finally the logged parasite density is reported with a mean value due to the normalising property of the log transformation.

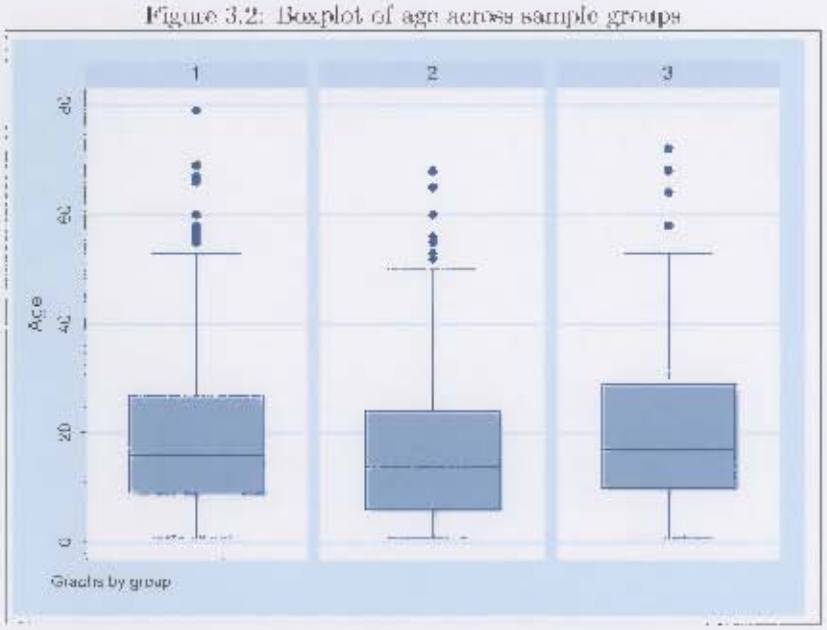
Table 3.2: Demographics of the sample

Demographics	Sample Groups					
	1 (none observed)		2 (excluded)		3 (included)	
Count	305		171		103	
Mutations	*Sensitive*	*Resistant*	*Sensitive*	*Resistant*	*Sensitive*	*Resistant*
	220 (89%)	27 (11%)	128 (84%)	24 (16%)	82 (88%)	11 (12%)
Treatment outcome	*Success*	*Failure*	*Success*	*Failure*	*Success*	*Failure*
	236 (86%)	40 (14%)	134 (81%)	32 (19%)	93 (91%)	9 (9%)
Gender	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
	139 (46%)	166 (54%)	88 (51%)	83 (49%)	50 (49%)	53 (51%)
Site	*Mpm*	*Moz*	*Mpm*	*Moz*	*Mpm*	*Moz*
	228 (75%)	76 (25%)	104 (62%)	65 (38%)	72 (70%)	31 (30%)
Age	*Median*	*Range*	*Median*	*Range*	*Median*	*Range*
	16	1-79	14	1-68	17	1-72
Parasite density	*Geom. mean*	*Range*	*Geom. mean*	*Range*	*Geom. mean*	*Range*
	22630	1003-33933	23861	1008-598000	31462	1040-332000
Log Parasite density	*Mean*	*Range*	*Mean*	*Range*	*Mean*	*Range*
	4.35	3-5.5	4.38	3-5.8	4.5	3-5.5
Parasite clearance time	*Median*	*Range*	*Median*	*Range*	*Median*	*Range*
	2	1-7	2	1-7	3	1-7

Significant differences across the three sample groups were found for site

(p=0.009) and age (p=0.03). A higher proportion of patients from Mpumulanga (and a lower proportion from Mozambique) were in group one, and vice versa for group two. Boxplot 3.2 suggests that patients in group two were slightly younger than the other groups. There was some evidence of a difference across treatment outcomes (p=0.06) with a slightly higher proportion of treatment failures in group two and slightly fewer failures in group three. Chi-squared tests (or Fisher's exact tests) were used for categorical variables and Kruskal Wallis tests were used for continuous variables.

Figure 3.2: Boxplot of age across sample groups



Drug resistance to sulfadoxine and pyrimethamine can be defined according to the number of mutations in the two enzymes dihydrofolate reductase (dhfr) and dihydropteroate synthetase (dhps) respectively. These two enzymes are in the malaria parasites' pathway for synthesizing folate, which is required for cell division. The parasite becomes increasingly resistant to the effects of SP as it accumulates mutations in these two enzymes. The 'full-house' or complete set of mutations is a triple mutation in dhfr together with a double mutation in dhps but patients could carry any number in between and resistance can be partial.

Amongst the 103 patients included in the analysis, there were 34 patients (36%) with infections that were classified as 'sensitive' because there were no dhfr or dhps mutations, 11 (12%) patients exhibited infections that were fully resistant to SP (five mutations), and 48 (52%) patients had mixed mutations. 90% of the infections classified as 'mixed' had a favourable treatment outcome, while all the sensitive infections and 55% of the resistant infections were treat-

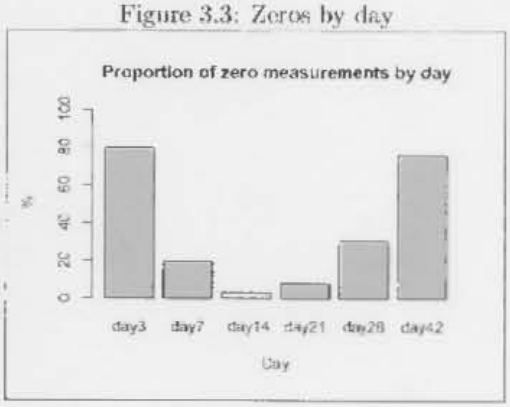
ment successes. The mutations variable used in the analysis (and displayed in table 3.2 above) was created as a full set of five mutations (resistant) versus the rest.

With regards to treatment outcome, there were 93 (90%) treatment successes, 1 (1%) patient lost to follow up, and only 9 (9%) treatment failures. The single loss to follow up was dropped when exploring the treatment outcome variable.

3.3.2 Zero gametocytes measurements

For the sample of 103 patients analysed, 35% of all the post day 0 gametocyte readings were zeros. Figure 3.3 below displays the proportion of gametocyte measurements that were zero by day.

Figure 3.3: Zeros by day



As expected, the lowest number of zeros were seen around day 14 which corresponds to the timeframe that the gametocyte prevalence and densities peak. The highest number of zeros were seen early and late in the course of the infection when gametocyte densities are lower and hence more likely to be close to the detectable limit.

3.3.3 Patient category

The three empirically created patient categories for the 103 patients included in the modeling were compared with regards to demographic and diagnostic variables in an attempt to see whether they characterised specific patient features.

Fishers Exact test was used to assess dependence between the categorical variables due to small numbers in the cells, and the Wilcoxin Rank Sum or Kruskal-Wallis tests were used to test for differences in continuous variables

across groups given the skew distributions. The Stata output for this section can be found in Appendix D.1.2 and the results are summarised in table 3.3 below.

Table 3.3: Univariate analysis of determinants of patient category

Variable		Patient Category			Pvalue	Test
		1	2	3		
Mutation	Sensitive	60 (73%)	9 (11%)	13 (16%)	0.30	1
	Resistant	7 (64%)	3 (27%)	1 (9%)		
Treatment Outcome	Success	66 (71%)	14 (15%)	13 (14%)	0.47	
	Failure	7 (78%)	0 (0%)	2 (22%)		
Site	MPM	50 (69%)	10 (14%)	12 (17%)	0.70	
	MOZ	24 (77%)	4 (13%)	3 (10%)		
Gender	Male	35 (70%)	7 (14%)	8 (16%)	0.95	
	Female	39 (74%)	7 (13%)	7 (13%)		
Parasite Clearance Time	Median	3	2	3	0.38	2
	IQ Range	2-3	2-7	3-7		
Age	Median	18	11	20	0.09	
	IQ Range	10-29	9-14	12-50		
Logged Parasite Density	Mean	4.5	4.5	4.6	0.37	
	Std Dev	0.5	0.5	0.5		

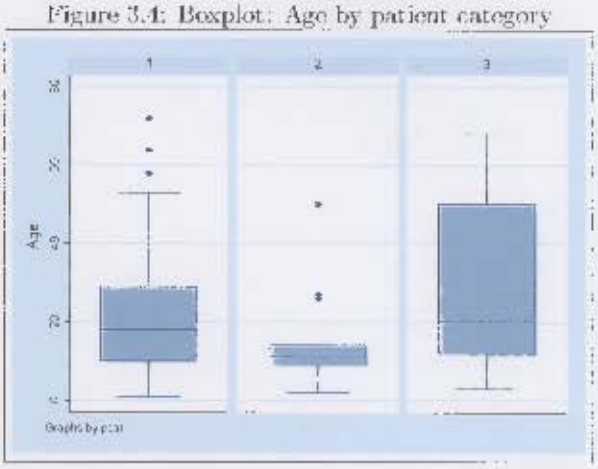
1 = Fisher's exact test; 2 = Kruskal Wallis Test
The percentages shown are row percentages

The patient categories were not significantly associated with mutations, whether all mixed strains were considered sensitive (p=0.3), or when mixed infections were grouped together seperately (0.68) . There was also no association between the patient categories and treatment outcome (p=0.47). It was surprising not to find any evidence of an association here but it is likely that this could be a power problem when considering the small number of patients that failed treatment (9) or were found to carry a fully resistant strain (11). No associations were found with site (p=0.70) or gender (p=0.95).

There was also no evidence of a difference across patient categories in median parasite clearance times (p=0.38) or logged parasite density levels (p=0.37). Examining age with the patient categories revealed that patient category two had the lowest median age with the least variation. Categories one and three had similar medians with category three also having more variation. A Kruskal Wallis test revealed slight evidence (p=0.09) for a true difference in the underlying age distribution. The boxplot is shown in figure 3.4:

A polytomous logistic regression was run in an attempt to get a multivariate view of possible associations with the patient categories (hence the patient categories variable was the response variable). The regression was run on each

Figure 3.4: Boxplot: Age by patient category



covariate (independent variable) separately as well as on all covariates together with either treatment outcome or mutation (since these 2 variables were highly correlated). This was done with patient category one being the base or reference category.

Only the mutation variable was found to have any significance (p=0.065) for category two relative to category one. The Relative Risk of 1.39 suggests that having a resistant strain makes one 1.39 times as likely as somebody with a sensitive strain to be in category two compared to category one. Note that this variable only became significant after adjusting for age and site. The treatment outcome variable does not reflect this due to the fact that there are no resistant outcomes in category two.

3.3.4 Exploring the covariates

The associations between the other covariates in the group analysed are summarised in table 3.4 below. For further details of this exploration refer to Appendix D.1.2.

Table 3.4: Exploring the association between covariates

Test	Variables		Pvalue
Fisher's exact test	Treatment outcome	Mutation	0.001 **
	Site	Mutation	0.09 *
	Site	Treatment outcome	0.44
	Site	Parasite clearance time	0.02**
Wilcoxin Rank Sum	Treatment outcome	Log parasite density	0.015**
	Mutation	Log parasite density	0.54
	Mutation	Age	0.21
	Site	Parasite clearance time	0.006**
	Site	Age	0.16
	Site	Log parasite density	0.42
	Treatment outcome	Age	0.33
Spearman Correlation	Age	Log parasite density	0.40
	Age	Parasite clearance time	0.17

* – significant at 10%; ** significant at 5%

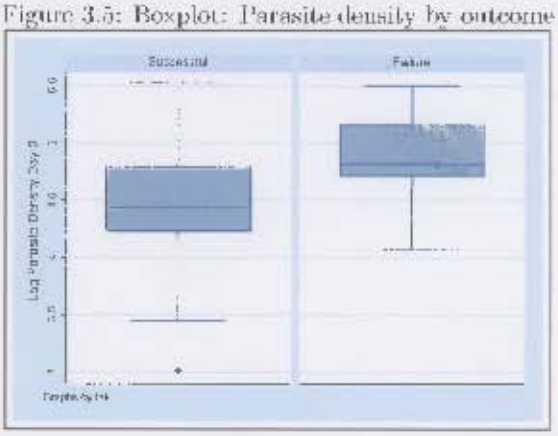
Treatment outcome and mutations were associated (p=0.001) as one would expect. Sensitive strains invariably resulted in successful outcomes (95%) whereas resistant strains had just over a 45% chance of failing to respond to treatment.

Slight evidence was found of an association between site and mutations (p=0.09) when conducting a one-tailed test. There was a higher proportion of patients with a resistant strain in Mozambique (20%) compared to Mpumulanga (8%) as expected. As chloroquine was the treatment policy in Mozambique at the time of the study, the resistance to SP can potentially be explained by the fact that the Mozambique sites border KwaZulu Natal where resistance to SP is rife as it was the first place to implement SP as a treatment policy. Drug pressure to SP has therefore been exerted in that area for longer than elsewhere resulting in the highest treatment failure rates in Africa being recorded there. (Bredenkamp et al., 2001)

Site and parasite clearance time were associated, whether parasite clearance time was considered as a categorical variable (p=0.02) or as a continuous variable (p=0.006). Mozambique appeared to have patients clearing parasites more rapidly than Mpumulanga. Patients from Mozambique also had a lower median parasite clearance time compared to Mpumulanga. This could reflect the differential degree of immunity between the two sites.

There was evidence of a relationship between treatment outcome and logged parasite density on day 0 with higher baseline asexual parasite levels appearing to be associated with treatment failures (p=0.015). Boxplot 3.5 depicts this association. Interestingly there was no association detected between mutation and logged parasite density at baseline (p=0.54).

Figure 3.5: Boxplot: Parasite density by outcome



There were no associations found between site and treatment outcome (p=0.44), age (p=0.16) or logpdens (p=0.42); or between age and treatment outcome (p=0.33) or mutation (p=0.21). Spearman correlation coefficients for age and parasite density and for age and parasite clearance time were respectively 0.08 and 0.14 and hence did not suggest much of a relationship.

3.3.5 Conclusions from data exploration

Apart from treatment outcome and mutation, there appears to be no concern regarding multicollinearity in the modeling. Mozambique is characterised by a slightly higher level of resistance to SP as well as a higher degree of immunity when compared to Mpumulanga.

The most salient point to emerge from the comparison of the group analysed with the groups excluded, was the fact that group 2 had more failures and a higher proportion of patients from Mozambique than the other groups. This makes sense as more resistance should lead to more treatment failures and failures are likely to be withdrawn before sufficient gametocyte readings have been taken.

The three empirically created patient groups do not clearly correspond to specific patient profiles as measured by the demographic and disease-specific covariates. These groupings seem to present the effect of characteristics not documented in this study (e.g. delay in seeking treatment) or may be a feature of the data sampling.

3.4 Nonlinear Mixed Effects Modelling

The modeling used in this dissertation was broken up into two broad phases. The first phase involved finding a nonlinear function that could potentially explain the underlying structure of the gametocyte density curves. The second phase examined the impact of the covariates on the patient density-time profiles. The modeling was conducted using R 2.2.0 (R Development Core Team, 2005) and in particular the *nlme* package (Pinheiro et al., 2006).

3.4.1 Finding the structure of the gametocyte density-time profiles

The strategy I used for this phase of modeling started with a nonlinear least squares fit that ignored the grouping in the data to get an idea of starting values ('nls' fit) . Then individual regressions were fitted per subject, in order to get an idea of the variability in the parameters as well as the extent of data sparcity i.e. by seeing how many patients did not have sufficient data to obtain estimates for a particular function ('nlsList' fit).

I then moved onto the mixed effects modeling ('nlme' fit) and started with the most general variance-covariance structure. If it could solve I examined the results for signs of possible improvement, for example if a particular random effect was really small relative to the estimated fixed effect then I would try dropping this. I then compared the original model to the reduced model with a likelihood ratio (LR) test and continued the process until I had the 'best' random effects structure for the model.

If the most general model could not solve, I simplified the structure and then compared various options to each other with LR tests. The results of the individual regressions ('nlsList' fits) were used to guide this simplification i.e. based on apparent variability in the individual parameter estimates.

The same process was conducted for each different dataset and each different function that was attempted. Note that the important use of the LR test in this strategy meant that I had to use Maximum Likelihood (ML) as opposed to Restricted Maximum Likelihood (REML) in the estimation process. The reason is that any change to either the fixed or the random effects structure affects the REML likelihood and hence different structures cannot be compared in this way if REML is used.

Note that due to the large number of models that needed to be attempted across the 12 datasets, a decision was made not to use alternative methods (such as bootstrapping) to form intervals for those models that could not calculate the standard errors directly.

Functions attempted

The results from trying to fit various different nonlinear functions to the data are briefly presented below. The estimates for the fixed effects and the random effects are shown in the tables together with the AIC's and values for the log likelihood functions. Note that a random effect that is labelled with two parameters (e.g. $\beta_0\beta_1$) refers to the correlation between those two random effects, whereas a random effect for a single parameter (e.g. β_1) is expressed as a standard deviation. The appropriate code can be found in appendix D.2.1.

The last two wave-type functions were only attempted on the two largest datasets i.e. 56 and 103 patients. This was due to the fact that all three patient groupings were required to investigate these forms and it makes little sense to attempt to fit a wave-type function after dropping those groups that exhibited oscillation.

One Compartment Model: $\beta_0 * \left(exp^{(-\beta_1 * X)} - exp^{(-\beta_2 * X)}\right)$

This is the basic one compartment model used in pharmacokinetic studies where a drug gets absorbed into the body and then eliminated. β_1 can be thought of as the elimination rate and β_2 as the absorption rate.

Data1 - zeros missing: for five of the six datasets the best random effects structure was a general matrix with random effects on β_0 and β_1. In general, the same pattern was observed in that difficulties were encountered (confidence intervals could not be estimated) when attempting to fit an unstructured matrix with random effects on all three parameters but the situation was stable with random effects only on β_0 and β_1. Furthermore the completely unstructured models that could solve, all estimated low variance for β_2's random effect and almost no correlation between the random effects for β_2 and the other parameters.

A block diagonal structure was found to be best for the biggest dataset of 103 patients. This involved allowing the random effects for β_0 and β_1 to correlate with β_2's random effect being independent. Under this scenario the estimated standard deviation for β_2's random effect was no longer negligable. Results are summarised in table 3.5.

Table 3.5: Data1 - One compartment model

Datasets		Fixed Effects			Random Effects				AIC	Loglk
Pcat	Size	β_0	β_1	β_2	β_0	β_1	β_2	$\beta_0\beta_1$		
1	34	11.76	0.018	0.40	2.98	0.011	N/A	0.74	617	-301
1	74	10.25	0.015	0.41	2.89	0.010	N/A	0.73	1096	-541
1+2	47	11.31	0.015	0.43	2.72	0.009	N/A	0.64	848	-417
1+2	88	10.20	0.014	0.43	2.73	0.009	N/A	0.65	1339	-662
1+2+3	56	10.65	0.012	0.47	2.73	0.009	N/A	0.65	1055	-520
1+2+3	103	10.77	0.016	0.33	3.00	0.009	0.126	0.65	1600	-792

Data2 - zeros changed: Only very limited mixed effects models (with a single random effect) could be solved for datasets 56, 74 and 103 but in all cases the nonlinear least squares (nls) fit was better i.e. when random effects are ignored and the same shape is fitted for all individuals. An unstructured matrix could be solved for dataset 34 but, as with the other datasets, the nls fit was better.

These one compartment models appeared to be extremely sensitive to starting values in the sense that there were convergence issues when the starting values were changed slightly.

Biexponential: $A_1 * exp^{(-exp^{(lrc_1)} * X)} + A_2 * exp^{(-exp^{(lrc_2)} * X)}$

The biexponential is similar to the previous function but has an extra scale parameter. Note the different paramaterisation here compared to the one compartment model above. This form just ensures that the rates lrc_1 and lrc_2 are positive. I considered this reparamerisation for the one compartment model but as all my estimates were positive anyway I opted to use the simpler form.

Data1 - zeros missing: Mixed effects models could only be fitted to the two smallest datasets corresponding to patients with at least four nonzero gametocyte readings and where patients in category three are excluded. In these cases the best random effects structure was diagonal with random effects on lrc_1 and lrc_2. Unstructured matrices would not solve. Results are summarised in table 3.6.

Table 3.6: Data1 - Biexponential model

Datasets		Fixed Effects				Random Effects		AIC	Loglk
Pcat	Size	A_1	Lrc_1	A_2	Lrc_2	Lrc_1	Lrc_2		
1	34	-21.58	-1.82	21.68	-3.20	0.30	0.25	636	-311
1+2	47	-20.43	-1.82	20.56	-3.28	0.29	0.26	912	-449

Data2 - zeros changed: Mixed effects models could only be solved for datasets

34 and 88 and again the best random effects structure was diagonal with random effects on lrc_1 and lrc_2 with unstructured matrices not being able to solve. In both cases the nonlinear least squares (nls) fit was better!

The negative correlation between the fixed effects A_1 and A_2 was almost perfect and hence this model appeared overparamaterised. Note that this situation is somewhat contrived due to the fact that all subjects begin with a zero response at day 0. Examining the function it is apparent that these two estimates need to cancel each other out when $X = 0$ in order to predict a starting point at zero. A Biexponential structure does not seem to be appropriate for this data and it was therefore not surprising that there were estimation difficulties. This is also confirmed by the lower AIC's for the one compartment compared to the Biexponential models within the different datasets.

Two Compartment model: $\left(A_1 * exp^{(-\beta_1 * X)}\right) + \left(A_2 * exp^{(-\beta_2 * X)}\right) + \left(A_3 * exp^{(-\beta_3 * X)}\right)$

This function is analagous to the two compartment model used in pharmakokinetic modeling when the absorption and elimination of a particular drug follows a more complex biological procedure than what can be captured with a one compartment model.

Not even one dataset could solve with nonlinear last squares i.e. without any random effects, and it was obvious that the data could not support this structure.

Critical exponential: $\left(A + (B + C * X) * R^X\right)$

The critical exponential curve is a variation of the double exponential (a special limiting case) and can take a variety of shapes.

Data1 - zeros missing: The number of individual regression fits that could be solved was virtually the same for the bigger datasets as the smaller ones i.e. could not fit the extra patients with one less datapoint. Limited mixed effects models could be fit to all of the datasets though the fits obtained for datasets 56 and 88 were worse than the nonlinear least squares fit that ignored random effects. A diagonal structure with random effects on C and R was found to be best for dataset 34 and dataset 74, with datasets 47 and 103 being best with a single random effect only on C and R respectively. Results are summarised in table 3.7.

Table 3.7: Data1 - Critical exponential model

Datasets		Fixed Effects				Random Effects		AIC	Loglk
Pcat	Size	A	B	C	R	C	R		
1	34	3.75	-3.6	2.24	0.90	0.75	0.022	642	-314
1	74	3.32	-3.22	1.96	0.90	0.66	0.019	1135	-560
1+2	47	4.58	-4.20	1.95	0.90	0.59	N/A	963	-475
1+2+3	103	5.04	-4.90	1.96	0.88	N/A	0.025	1772	-880

Data2 - zeros changed: For four of the six datasets the best structure was found to be an unstructured matrix with random effects on C and R. A different structure with a random effect on C only was found for datasets 56 and 103. In no cases could a completely unstructured matrix for all parameters be estimated. Furthermore the fixed effects for the A and B parameters were not significant in any models. High correlations with these two parameters also suggest over parameterisation. The A and B parameters were also highly correlated for the same reason as the biexponential function above. Results are summarised in table 3.8.

Table 3.8: Data2 - Critical exponential model

Datasets		Fixed Effects				Random Effects			AIC	Loglk
Pcat	Size	A	B	C	R	C	R	CR		
1	34	-0.65	0.60	2.09	0.93	0.61	0.013	-0.85	915	-449
1	74	-0.36	0.32	1.79	0.93	0.59	0.014	-0.78	1891	-937
1+2	47	-0.96	1.00	1.96	0.93	0.54	0.012	-0.82	1294	-639
1+2	88	-0.53	0.54	1.76	0.93	0.54	0.013	-0.75	2301	-1142
1+2+3	56	-0.17	0.41	1.73	0.93	0.30	N/A	N/A	1646	-817
1+2+3	103	0.04	0.19	1.54	0.93	0.33	N/A	N/A	2877	-1432

Modified Critical exponential: $(C * X) * (R^X)$

As neither A nor B was significant in the critical exponential models fitted above, a modified version of the equation without A or B was attempted next.

Data1 - zeros missing: The best random effects structure across all data sets was found to be an unstructured matrix for both parameters. Results are summarised in table 3.9.

Table 3.9: Data1 - Modified critical exponential model

Datasets		Fixed Effects		Random Effects			AIC	Loglk
Pcat	Size	C	R	C	R	CR		
1	34	2.29	0.922	0.75	0.017	-0.85	683	-335
1	74	2.03	0.921	0.70	0.016	-0.79	1207	-597
1+2	47	2.14	0.926	0.66	0.015	-0.81	989	-488
1+2	88	1.97	0.924	0.63	0.015	-0.75	1543	-765
1+2+3	56	2.02	0.929	0.68	0.016	-0.83	1210	-599
1+2+3	103	1.86	0.927	0.64	0.016	-0.79	1869	-928

Data2 - zeros changed: The best random effects structure was again found to be an unstructured matrix for both parameters. Results are displayed in table 3.10.

Table 3.10: Data2 - Modified critical exponential model

Datasets		Fixed Effects		Random Effects			AIC	Loglk
Pcat	Size	C	R	C	R	CR		
1	34	2.08	0.926	0.63	0.014	-0.84	906	-447
1	74	1.78	0.926	0.6	0.014	-0.77	1880	-934
1+2	47	1.96	0.929	0.56	0.013	-0.81	1279	-633
1+2	88	1.76	0.928	0.55	0.014	-0.73	2283	-1135
1+2+3	56	1.88	0.931	0.62	0.015	-0.83	1571	-779
1+2+3	103	1.68	0.930	0.57	0.015	-0.77	2776	-1382

It was noticeable that the individual regressions fitted for all subjects probably due to the fact that this function only has two parameters. The values for R seemed very robust and changed slightly for C.

Fourier: $A + B * sin\left(2\pi * \frac{X-E}{W}\right)$

The Fourier curve can be used to model data that exhibit cyclical or periodic behaviour i.e. the responses fluctuate up and down at regular time intervals.

Data1 - zeros missing: Mixed effects models could not solve for dataset 103 but could solve for dataset 56, where all three empirical patient categories were included but only patients with at least four nonzero readings were used. The best structure found was for a random effect on A only but the nonlinear least squares fit was better. The Fourier model did not appear to fit this data well, when it could solve a mixed effects model the estimated random effects were small i.e. a similar curve was fitted for everyone.

Data2 - zeros changed: The best random effects structure was a general unstructured matrix for dataset 56 and a general unstructured matrix with

A, B and E for dataset 103. Note that the mixed effects model fit was only marginally better than the nonlinear least squares fit for dataset 56 and that confidence intervals could not be estimated for any of these models. Results are shown in table 3.11.

Table 3.11: Data2 - Fourier model

Datasets		Fixed Effects				Random Effects				AIC	Loglk
Pcat	Size	A	B	E	W	A	B	E	W		
1+2+3	56	6.16	4.20	6.35	39.77	1.00	0.93	1.51	2.76	1848.3	-909.1
1+2+3	103	5.42	3.72	6.17	39.44	1.03	0.87	0.68	N/A	3194.9	-1586.4

The correlation matrices (tables A.1 and A.2 in appendix A) suggest that this structure is overparamaterised. It is possible that the Fourier function is capable of describing a wave-like pattern but that there is not enough data available here to distinguish between some of the parameters.

Double Fourier: $A + B * sin\left(2\pi * \frac{X-E}{W}\right) + G * sin\left(4\pi * \frac{X-P}{W}\right)$

This is an extension of the Fourier model and allows for a second wave in the data to be modelled before the first wave has declined to the same level that it started at.

Data1 - zeros missing: For dataset 56 the best structure was block diagonal with A & B together and E alone (note though that there was perfect correlation between A and B), while for dataset 103 the full unstructured matrix was best. Note that various random effects structures could solve for dataset 103. No confidence intervals could be estimated for any of these models. Results are shown in table 3.12.

Table 3.12: Data1 - Double fourier model

Datasets		Fixed Effects						Random Effects						AIC	Loglk
Pcat	Size	A	B	E	W	G	P	A	B	E	W	G	P		
1+2+3	56	6.39	4.33	6.25	37.91	2.83	2.70	1.21	0.93	0.96	N/A	N/A	N/A	1222	-600
1+2+3	103	5.73	4.07	6.33	38.17	2.71	2.90	1.29	1.04	1.35	1.84	0.55	1.19	1828	-886

The correlation matrix (table A.3 in Appendix A) displays signs of overparamaterisation with correlation coefficients above 0.90 for three different parameter pairs.

Data2 - zeros changed: the full unstructured matrix was best for both datasets. Confidence intervals could not be estimated for most of the models that were tried. Table 3.13 summarises these results.

Table 3.13: Data2 - Double fourier model

Datasets		Fixed Effects						Random Effects						AIC	Loglk
Pcat	Size	A	B	E	W	G	P	A	B	E	W	G	P		
1+2+3	56	5.97	4.71	6.41	38.61	2.72	2.83	1.21	1.33	1.65	1.86	0.58	1.29	1581.5	-762.8
1+2+3	103	5.25	4.26	6.24	38.50	2.20	2.83	1.24	1.23	1.74	1.62	0.74	1.12	2796.4	-1370.2

Again the correlation matrices (tables A.4 and A.5 in Appendix A) suggest that this structure is overparamaterised.

Despite the clear problems with estimating this structure, the fitted plot (plot C.2 in appendix C) appears to be promising. Further investigation with a richer dataset is warranted.

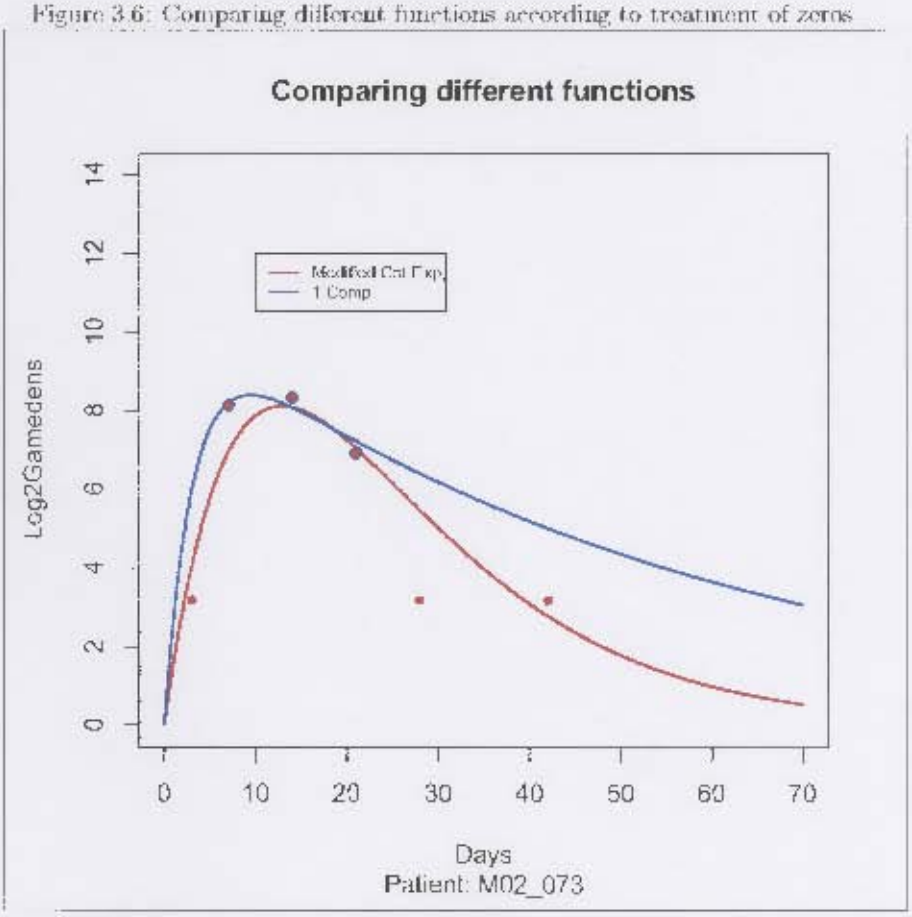
Final structure

The modified critical exponential function is the one that has been chosen as most appropriate for this data. This decision is based on the clear stability that this structure has demonstrated across the various datasets and data sizes, particularly with regards to the parameter R. The fitted plot C.1 in appendix C shows that the function appears to be capable of taking on a wide array of shapes with just two parameters.

It should be noted that the AIC's from the one compartment model were better than those from the modified critical exponential function for Data1, but unfortunately the one compartment model did not perform well with Data2. In Data1 the zeros were effectively dropped thereby allowing a smoother function to be fitted as the fitted curve does not need to reach down to the extremely small observations. When the zeros are replaced by a small number (in this case 8), the function then attempts to take these dips into account.

Figure 3.6 below shows the fitted curves for a particular patient for firstly the one compartment model applied to Data1, and secondly for the chosen modified critical exponential function applied to Data2. The three red points on the same horizontal band therefore refer to the three zeros for this patient and hence are ignored in the one compartment model but used in the modified critical exponential function. This clearly shows how using Data1 leads to a much flatter right-hand tail corresponding to a very slow elimination of gametocytes that does not appear to be biologically plausible. Therefore the one compartment model could not be chosen despite having lower AIC's for the datasets within Data1.

Figure 3.6: Comparing different functions according to treatment of zeros



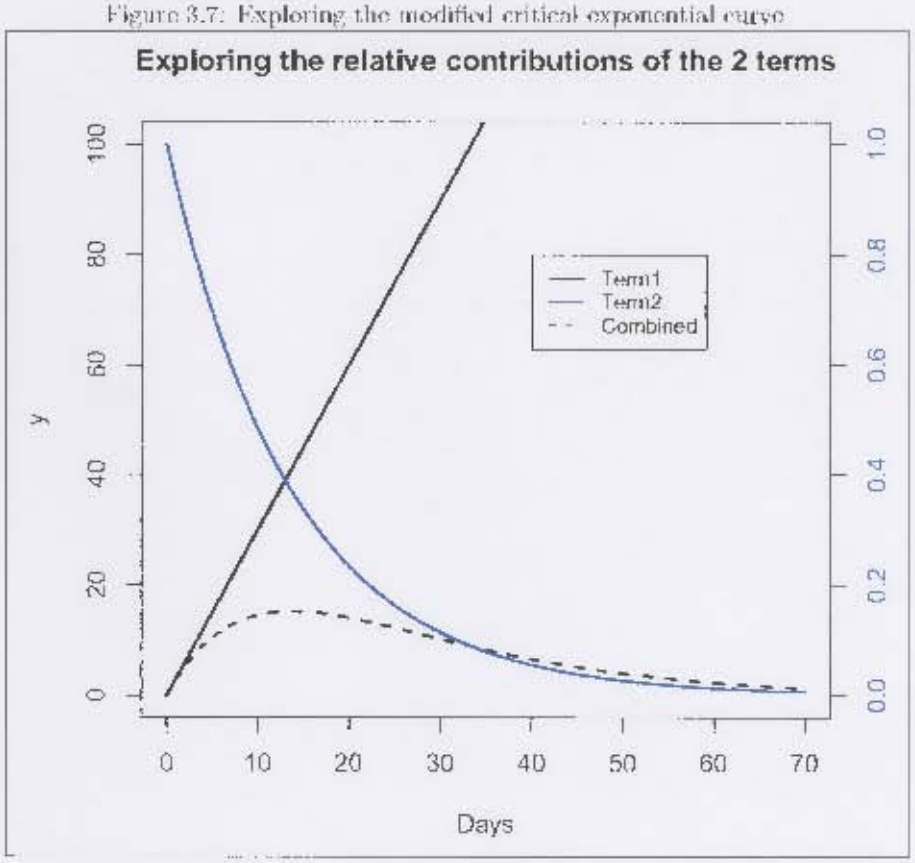
The critical exponential function also produced better AIC's than the modified form for Data1, yet that function was clearly overparameterised with respect to the A and B parameters.

The modified critical exponential function is similar to an ordinary exponential decay curve: $\beta_0 * exp^{(-\beta_1 X)}$. This is apparent when it is rewritten as $(C * X) \times exp^{(X log R)}$ where $logR$ can be substituted for β_1, note though that this constrains R to fall between zero and one in order to give a negative value. The difference is that the first term $(C * X)$ is not constant like β_0 but rather depends on time. The ordinary exponential decay curve has the property that the ratio of the response at successive time points $(\frac{f(X+1)}{f(X)})$ is equal across the range of the function. This does not hold for the modified critical exponential function.

I explored the modified critical exponential function that was chosen (the

code for this can be found in appendix D.2.1) by examining the relative contributions of the two terms in the function, and the effect on the shape of the curve of varying one parameter while leaving the other constant. The relative contributions of the two terms in the final function are plotted below (plot 3.7). Note that the second term is plotted against the Y axis on the right that is on a different scale to the Y axis on the left.

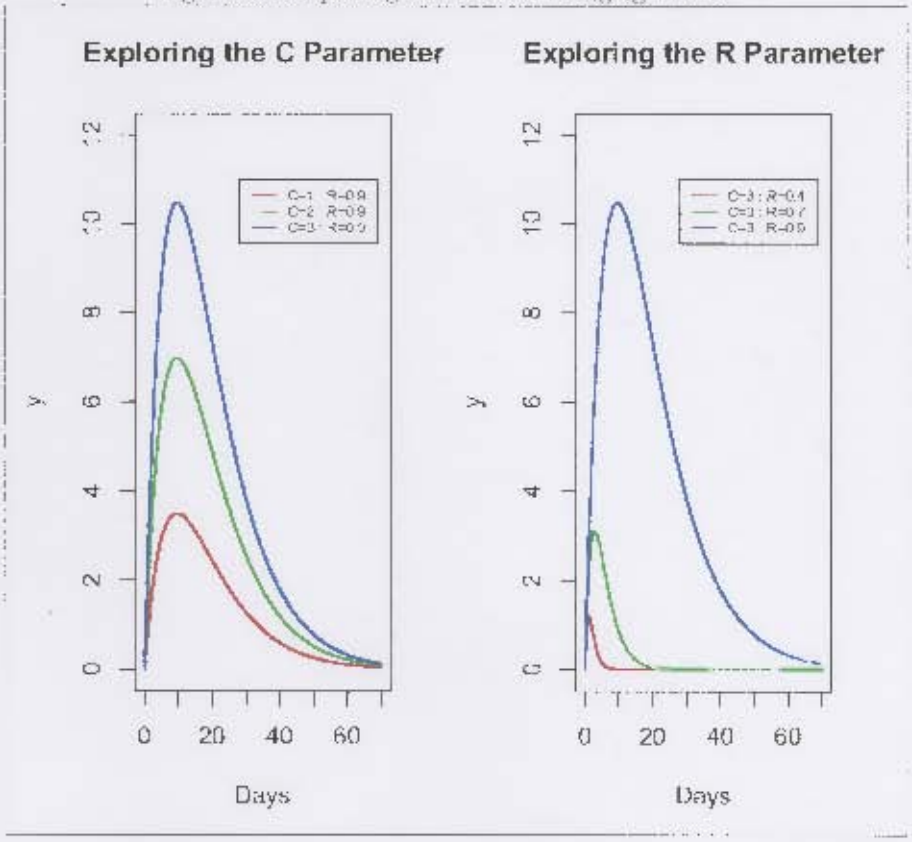
Figure 3.7: Exploring the modified critical exponential curve



The first term $(C * X)$ increases linearly with x. The contribution of the second term $exp^{X log R}$ is to exponentially decay the linearly increasing amount from the first term, with higher values for R leading to a slower rate of decay. The fact that term one increases with X is what differentiates this function from an ordinary exponential decay curve and allows the function to increase initially.

The effect of changing either the C or the R parameter is displayed in plot 3.8 below.

Figure 3.8: Exploring the effect of changing C & R



Changing C appears to affect the overall scale as well as both the rate at which the gametocytes increase initially, and the rate at which the gametocytes decrease from the peak. Changing R, on the other hand, appears to effect the timing of the peak as well as the overall scale. Higher values for C lead to larger rates of increase and decrease while higher values of R lead to the gametocytes peaking later on and to higher gametocyte densities (as higher values for R lead to a slower rate of exponential decay). Therefore both C and R together determine the shape of the curve. Since the chosen function is multiplicative in nature, it makes sense that the two parameters are inter-related and cannot be seperated into say a scale and a shape parameter.

3.4.2 Modelling covariates

The code for the second phase of modeling described below can be found in Appendix D.2.2. The initial model building phase focused on finding an appro

priate underlying structure. Based on all of the information presented above, the modified critical exponential structure was identified as being the most appropriate. This structure was therefore applied to the full dataset of 103 patients that included all 3 empirical patient categories and patients with at least 3 nonzero gametocyte readings.

This phase examines if any covariates can improve the model and explain some of the individual variation in the model parameters thereby reducing the random effects. As suggested by Pinheiro and Bates (2002), the strategy is to fit a model without any covariates and then plot the random effects from this model against the potential covariates that are candidates to enter the model. Any systematic pattern in the relationship between a potential covariate and a particular random effect would suggest that the particular covariate be included in the model.

The plots displayed in appendix B show the results of plotting the random effects from the appropriate model against the potential covariates for the two parameters C and R. Note that all covariates are included in these plots regardless of whether or not they are in the model. Any observed pattern between the random effects and a particular covariate should disappear after that variable is included in the model.

Covariates are added to the model one at a time starting with the one with the most obvious pattern. The new model is fitted and the process repeated with the new set of random effects. Wald type tests are used to test the significance of any fixed effect associated with a covariate. The model building process is first outlined in detail and then followed by tables that summarise the results.

Tables 3.14 and 3.15 below display the parameter estimates and model fit statistics at various stages of the model building process. In particular they show the stability of the individual parameter estimates regardless of the addition of extra covariates to the models as well as the decreasing values for the random effects, AIC, and log likelihood as new covariates enter the models.

The notation used to label the parameters is consistent with the notation used in the R output. Therefore C.age refers to the coefficient for age which would get used in the 2nd stage of the model to explain individual variation around the population average of the C parameter. Consequently C.intercept and R.intercept refer to the average values of the C and R parameters when all covariates equal zero and hence these should not be directly interpreted. Note that continuous covariates like age could be centered thereby making interpretation of the covariate pattern at zero more meaningful.

Data1 - zeros regarded as missing data

The results of this model building process are summarised in table 3.14 below.

No Covariates

The plot of random effects for C (B.1) suggested that patient category, site, treatment outcome and logged parasite density were the strongest candidates for entering the model and explaining some of the variation in the C parameter. The plot of random effects for R (B.2) showed a strong relationship with age followed by patient category, treatment outcome and site. It was interesting that the mutation variable did not appear to play a role.

First Covariate - Age

Age was the first covariate introduced to the model based primarly on the plot of the random effects for R. Likelihood ratio tests (LR) tests confirmed that the model with age on both parameters was better than a model with age on R only (p=0.029). The model with age on both C and R was an improvement over the model with no covariates (p=0.0001). Age was significant on both C (p=0.0322) and R (p=0.0001), and these 2 terms were also found to be jointly significant (p=0.0001).

The plot of random effects from the model with age in it for C (B.3) suggested that patient category, logged parasite density, site, and treatment outcome were the strongest candidates for entry. The plot of random effects for R (B.4) also suggested patient category, treatment outcome and site.

At this stage treatment outcome looked quite promising and it can be thought of as a proxy for resistance and hence was expected to play an important role. A model with the treatment outcome variable was therefore fitted. A model with treatment outcome only on the R parameter was no worse than one with treatment outcome on both C and R (p=0.16). However treatment outcome was not significant in the model (p=0.68) and the model including treatment outcome was also no better than the model from the previous step with age only (p=0.71). This somewhat unexpected result could be related to power seeing as there were only 9 patients that failed treatment.

Second Covariate - Site

Since including treatment outcome did not improve the model, the next covariate included was site. LR tests confirmed that the model with the covariate site on both parameters was better than a model with site on R only (p=0.0018) or a model with site on C only (p=0.0001). Including site on both parameters also led to a model that was an improvement over the model with no covariates (p=0.0006). Site was significant on both C (p=0.0018) and R (p=0.0001), and these two terms were also found to be jointly significant (p=0.0003).

The plot of new random effects for C (B.5) suggested that patient category

and logged parasite density were the strongest candidates to improve the model. The plot for R (B.6) suggested that only patient category could potentially improve the model.

Third Covariate - Logged parasite density
LR tests confirmed that the model with logged parasite density on both C and R was no better than a model with logged parasite density on C only (p=0.46) and that the model with logged parasite density on C only was slightly better than the model with logged parasite density on R only (they had the same degrees of freedom and hence no pvalue but a marginally lower AIC). This model was an improvement over the model with no covariates (p=0.0046) and the logged parasite density variable was highly significant (p=0.0044).

The plot of random effects for C (B.7) and for R (B.8) now only suggested that perhaps patient category could improve the model. I also investigated including parasite clearance time both as a continuous variable and as an ordered factor but it did not improve the model.

Fourth Covariate - Patient category
LR tests confirmed that the model with patient category on both C and R was better than a model with patient category on either C only (p<0.0001) or R only (p=0.0018). This model was an improvement over the previous model (p<0.0001). As the plots suggest, the patient category variable on the C parameter was significant (p=0.0007) for category three vs category one but not for category two (p=0.27) vs category one. With respect to the R parameter, patient category was significant when comparing either category two (p=0.0012) or category three (p<0.0001) to the reference category i.e. category one. Plots (B.9) and (B.10) no longer suggested any other candidate covariates.

Several things were attempted in order to try and improve the model. Firstly, as age was now marginally significant (p=0.077) on the C parameter, this model was compared to a model where age was dropped from the C parameter (but still left on R). The LR test suggested a marginal improvement (p=0.06) to the model when age was included on both parameters and so age was left on C.

A possible interaction between site and mutation was also investigated by fitting another model with mutation as a main effect together with its interaction term with site. The model with mutation and its interaction with site in it was no improvement (p=0.46) over the current model.

The final model resulted in the random effects for C and R being reduced by 16.8% and 35.2% respectively. The AIC decreased by 2.6% and the Log Likelihood by 3.6%.

The Diagnostic plots did not raise any alarm bells with regards to violations of the assumptions. There did not appear to be any serious departures from

normality for the random effects on C and R (plot B.11) while plot B.12 does not suggest non-constant variance. Lastly plot B.13 shows a fairly good level of agreement between the fitted and observed values across the whole range of fitted values.

Table 3.14: Data1 - Model building

		Variable entering model				
Parameters		**None**	**Age**	**Site**	**Pdens**	**Pcat**
Fixed Effects	**C. Intercept**	1.86	2.05	2.35	1.32	1.31
	C.age		-0.009	-0.008 *	-0.008	-0.007*
	C.site (MPM)			-0.47	-0.47	-0.41
	C.logpdens				0.23	0.25
	C.pcat2					-0.2**
	C.pcat3					-0.6
	R. Intercept	0.927	0.918	0.909	0.909	0.906
	R.age		0.0004	0.0004	0.0004	0.0003
	R.site(MPM)			0.015	0.015	0.013
	R.pcat2					0.013
	R.pcat3					0.018
Random Effects	**C**	0.638	0.629	0.587	0.581	0.53
	R	0.016	0.014	0.013	0.013	0.01
	Corr (C,R)	-0.79	-0.80	-0.76	-0.78	-0.74
Whole model	**AIC**	1870	1858	1847	1841	1821
	Log Lk	-929	-921	-914	-910	-896
	Residual error	1.118	1.12	1.12	1.12	1.12

All parameters significant at 5% except where indicated otherwise
* significant at 10%, ** not significant

Data2 - zeros changed to the lower limit of detection

The results of this model building process are summarised in table 3.15 below.

No Covariates

The plot of random effects for C (B.14) suggested that patient category and site were the strongest candidates for entry. The plot for R (B.15) indicated age followed by patient category, site and treatment outcome. Again the mutation variable did not appear to play a role.

First Covariate - Age

Age was the first covariate introduced to the model based primarly on the plot for R. As the plot suggest, LR tests confirmed that the model with age on R only was preferred to a model with age on both parameters (p=0.10), and that this model was an improvement over the model with no covariates (p=0.003). Age was significant on R (p=0.0016).

The plot of random effects for C (B.16) suggested that patient category and site were the strongest candidates to improve the model. There was some sign that treatment outcome could maybe play a role. The plot for R (B.17) also suggested patient category and site as the main candidates with treatment outcome showing some promise. Note that both these plots also suggested a slight nonlinear relationship with age even though it was already in the model. I considered modeling age in a nonlinear way through polynomials but decided that the apparent pattern was not strong enough to warrant this.

Second Covariate - Site

LR tests confirmed that the model with site on both parameters was better than a model with site on R only (p=0.0004) or C only (p<0.0001). This model was an improvement over the model with no covariates (p<0.0001). Site was significant on both C (p=0.0001) and R (p<0.0001), and these 2 terms were also found to be jointly significant (p<0.0001).

The plot of random effects for C (B.18) suggested that patient category and logged parasite density were the strongest candidates to improve the model. The plot for R (B.19) suggested that patient category and possibly logged parasite density and treatment outcome could potentially improve the model.

Third Covariate - Logged parasite density

LR tests confirmed that the model with logged parasite density on both C and R was no better than a model with logged parasite density on R only (p=0.47) and that the model with logged parasite density on R only was slightly better than the model with logged parasite density on C only (they had the same degrees of freedom and hence no pvalue). The model with logged parasite density on C only was an improvement over the previous model (p=0.0012) and the logged parasite density variable was highly significant (p=0.0009).

The plot of random effects for C (B.20) suggested that patient category could improve the model with slight signs for age and logged parasite density on the C parameter (already on R). The plot for R (B.21) suggested patient category and maybe treatment outcome as covariates that could lead to an improvement.

Fourth Covariate - Patient category

Seeing as plot B.20 suggested a possible relationship between the random effects for C and age, the effect of adding age on to the C parameter was checked but did not improve the model (p=0.21). The same result ocurred for attempting logged parasite density (p=0.47). The effect of adding the treatment outcome variable to R was also assessed but found not to improve the model (p=0.37).

LR tests confirmed that the model with patient category on both C and R was better than a model with patient category on either C only ($p<0.0001$) or R only (p=0.011). This model was an improvement over the previous model (p=0.0001). As the plots suggest, the patient category variable on the C parameter was significant (p=0.002) for category three vs category one but not for category two vs category one (p=0.72). With respect to the R parameter, patient category was significant when comparing either category two (p=0.0005) or category three (p=0.0007) to the reference category i.e. category one. Plots (B.22) and (B.23) no longer suggest any other candidate covariates.

Several things were again attempted in order to try and improve the model. Firstly the mutation variable was added but did not improve the model (p=0.45). A possible interaction between site and mutation was again investigated but the model with this interaction in it had no improvement (p=0.30) over the model without the interaction. I again investigated including parasite clearance time both as a continuous variable and as an ordered factor but it did not improve the model in either case.

The final model resulted in the random effects for C and R being reduced by 14.9% and 31.2% respectively. The AIC decreased by 1.8% and the Log Likelihood by 2.4%.

The Diagnostic plots did not raise any alarm bells with regards to violations of the assumptions. There did not appear to be any serious departures from normality for the random effects on C and R (plot B.24) while plot B.25 does not suggest non-constant variance and plot B.26 shows a fairly good level of agreement between the fitted and observed values across the whole range of fitted values. However there is a distinct line of values in both plot B.25 and plot B.26 that warrants comment. These values come from the zeros that were changed to be 8's and then when logged as (x+1) to the base two became 3.17. Plot B.26 also suggests that the model is not doing that well when it comes to fitting these psuedo-data points, especially for the larger fitted values.

Table 3.15: Data2 - Model building

Parameters		Variable entering model				
		None	Age	Site	Pdens	Pcat
Fixed Effects	C. Intercept	1.68	1.67	2.06	2.06	2.11
	C.site (MPM)			-0.54	-0.54	-0.50
	C.pcat2					-0.06**
	C.pcat3					-0.49
	R. Intercept	0.930	0.926	0.914	0.887	0.885
	R.age		0.0002	0.0002	0.0002	0.0002
	R.site(MPM)			0.017	0.016	0.015
	R.logpdens				0.006	0.006
	R.pcat2					0.010
	R.pcat3					0.013
Random Effects	C	0.57	0.57	0.52	0.52	0.49
	R	0.015	0.014	0.012	0.012	0.010
	Corr (C,R)	-0.77	-0.77	-0.71	-0.75	-0.75
Whole model	AIC	2776	2769	2751	2742	2727
	Log Lk	-1382	-1378	-1366	-1361	-1350
	Residual error	1.367	1.367	1.364	1.365	1.37

All parameters significant at 5% except where indicated otherwise
* significant at 10%, ** not significant

3.4.3 Final model

Table 3.16 below shows the results of firstly finding the best model for Data1 (under Model 1) and then fitting the same structure to Data2. Similarly, Model 2 is the best model structure found for Data2 and then it is applied to Data1 as well. The R output for the final models can be found in Appendix D.2.4.

The purpose of doing this was essentially to check the stability of the estimates when applied to a slightly different dataset. The table shows that there were not really any qualitative differences across the datasets. The inclusion of age appears doubtful as it is significant at 10% under Model 1 - Data1 but insignificant when applied to Data2 and also found not to be necessary for Model 2.

The 'best' model that will be used for further exploration and analysis is Model 2. The primary reason for this is that fact that Data2 is a richer dataset as the zeros were retained as opposed to simply being dropped. Figure 3.6 above shows how Data2 leads to more biologically plausible right-hand tails. The fact that the one compartment model could not fit Data2 was also key to the selection of the modified critical exponential function. Model 2 differs from Model 1 in that it does not include age on the C parameter and has logged parasite density on the R parameter rather than the C parameter.

Table 3.16: Final Models - checking stability

Parameters		Model 1		Model 2	
		Data1	**Data2**	**Data1**	**Data2**
Fixed Effects	**C. Intercept**	1.311	1.032	2.27	2.108
	C.Age	-0.007*	-0.004**	**NA**	**NA**
	C.site (MPM)	-0.415	-0.494	-0.429	-0.499
	C.logpdens	0.247	0.258	**NA**	**NA**
	C.pcat2	-0.200**	-0.090**	-0.148**	-0.059**
	C.pcat3	-0.602	-0.507	-0.604	-0.491
	R. Intercept	0.906	0.91	0.885	0.885
	R.age	0.0003	0.0003	0.0002	0.0002
	R.site(MPM)	0.013	0.015	0.014	0.015
	R.logpdens	**NA**	**NA**	0.005	0.006
	R.pcat2	0.018	0.011	0.012	0.011
	R.pcat3	0.013	0.014	0.018	0.013
Random Effects	**C**	0.5307	0.483	0.541	0.485
	R	0.0105	0.0101	0.0104	0.01
	Corr (C,R)	-0.74	-0.744	-0.745	-0.745
Whole model	**AIC**	1821.1	2728.4	1823.2	2727.1
	Log Lk	-895.5	-1349.2	-897.6	-1349.5
	Residual error	1.12	1.37	1.12	1.37

All parameters significant at 5% except where indicated otherwise
* significant at 10%, ** not significant

Referring to chapter 2, the first stage (within-subject) of the model for individual i can therefore be written as:

$$\boldsymbol{y_i} = \boldsymbol{f(x_j, \beta_i)} + \boldsymbol{e_i}$$

where: $\boldsymbol{f(x_j, \beta_i)} = (C_i * X_j) * (R_i^{X_j})$

and the second stage of the model can be written as:

$$\begin{aligned} C_i &= C.intercept + C.Site * site + C.pcat2 * pcat2 + C.pcat3 * pcat3 + c_i \\ R_i &= R.intercept + R.Site * site + R.age * age + R.logpdens * logpdens + \\ &\quad R.pcat2 * pcat2 + R.pcat3 * pcat3 + r_i \end{aligned}$$

where the notation for the parameters is consistent with the labelling above and in the R output, and the covariate labels are consistent with that given in the description of the covariates i.e. site refers to the site variable that takes on a zero (if from Mpumulanga) or a one (if from Mozambique).

So C_i refers to the value of the C parameter for the ith subject and consists of the group average C.intercept, plus a site effect if from Mozambique, plus

an effect of patient category (where pcat 2/3 refers to dummy variables that indicate the relevant category), plus some random deviation from this average c_i. R_i can be interpreted similarly but also includes terms indicating positive associations with age and logged parasite density at day 0.

Note that heterogeneous variance did not appear to be a problem with this data and so there was no need to introduce a variance function. The within-subject correlation was dealt with by including the subject-specific random effects on the C and R parameters.

It was noticeable that the empirical patient category variable was an important covariate even though logging the response did smooth the distinct shapes seen in plot 3.1.

Parameter interpretation

As there was no clear biological basis for explaining the different patient categories and shapes, a decision was made to refit the model without this variable. I also checked that no other covariates entered this revised model.

The results (along with 95% confidence intervals) for both the full and the revised models are summarised in table 3.17 below (and the output can be found in Appendix D.2.4). It was interesting that excluding the patient categories did not lead to any qualitative differences in the parameter interpretations and the main benefit of including patient category was to reduce the random variation in both the C and R parameters (from 0.522 to 0.485; and from 0.012 to 0.010 respectively).

Table 3.17: The Final Models - with confidence intervals

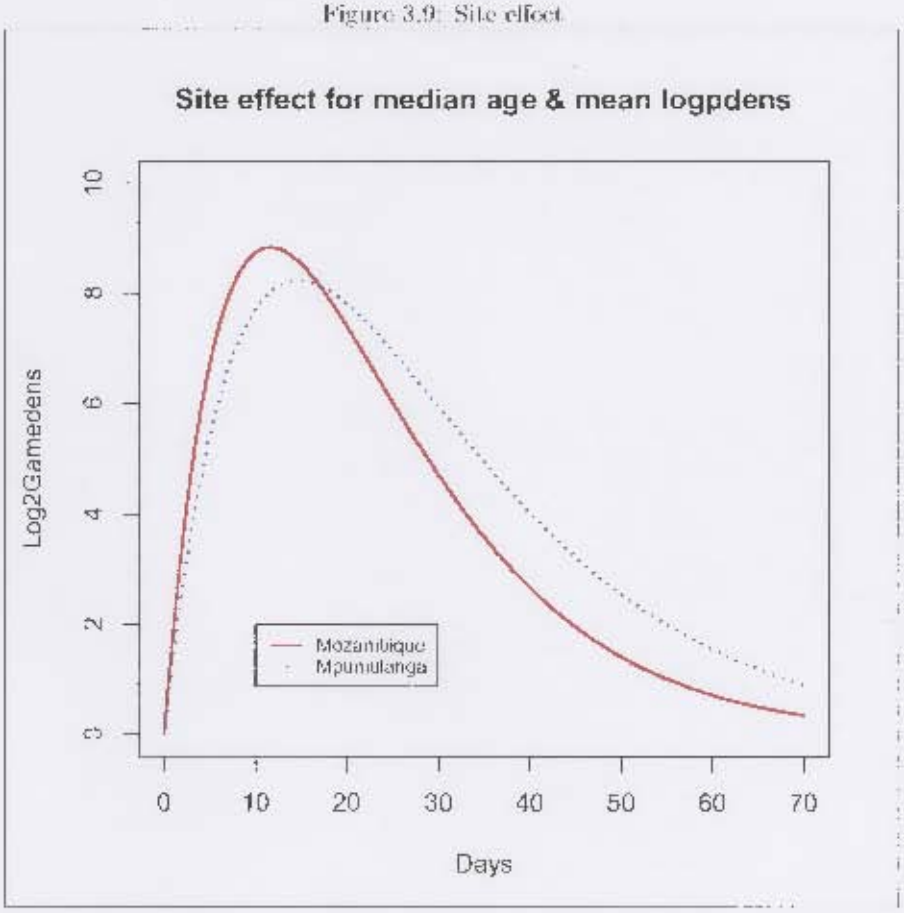
Parameters		Full Model	(95% CI)	Revised Model	(95% CI)
Fixed Effects	**C. Intercept**	2.108	(1.885 : 2.330)	2.059	(1.835 : 2.284)
	C.site (MPM)	-0.4999	(-0.750 : -0.247)	-0.541	(-0.804 : -0.277)
	C.pcat2	-0.059	(-0.381 : 0.263)	NA	NA
	C.pcat3	-0.491	(-0.799 : -0.183)	NA	NA
	R. Intercept	0.885	(0.868 : 0.901)	0.887	(0.869 : 0.904)
	R.site(MPM)	0.013	(0.009 : 0.021)	0.016	(0.010 : 0.023)
	R.age	0.0002	(0.0001 : 0.0003)	0.00015	(0.0000 : 0.0003)
	R.logpdens	0.006	(0.003 : 0.010)	0.006	(0.003 : 0.010)
	R.pcat2	0.010	(0.003 : 0.018)	NA	NA
	R.pcat3	0.013	(0.006 : 0.021)	NA	NA
Random Effects	**C**	0.485	(0.396 : 0.595)	0.522	(0.432 : 0.632)
	R	0.010	(0.008 : 0.013)	0.012	(0.009 : 0.015)
	Corr (C,R)	-0.75	(-0.848 : -0.587)	-0.75	(-0.846 : -0.609)
Whole model	**AIC**	2727.1	NA	2742.3	NA
	Log Lk	-1349.5	NA	-1361.2	NA
	Residual error	1.370	(1.286 : 1.460)	1.36	(1.282 : 1.453)

Data exploration revealed that mean logged asexual parasite densities at baseline were very similar for the two sites (4.50 & 4.48) but that the median age was slightly higher in Mpumulanga than in Mozambique (18 vs 17). The curves below are therefore plotted for site-specific median values of age and overall average logged asexual parasite density at baseline. Also note that the curves depicting the effect of site, age, and logged parasite density are generated from the reduced model without the patient category variable, whereas figure 3.12 depicts the effect of different patient categories and hence is generated from the full model.

The effects of the various covariates are as follows:

- Site: A patient from Mozambique would be expected to have a value for the C parameter that is approximately 0.5 units lower than a patient from Mpumulanga, and a value for the R parameter that is approximately 0.01 units higher, holding all else constant. The combined effect of this is displayed in plot 3.9 below and reveals that gametocyte carriage in patients from Mozambique tend to peak a little bit earlier, at a slightly higher peak, and then clear their gametocytes at a faster rate then patients from Mpumulanga.

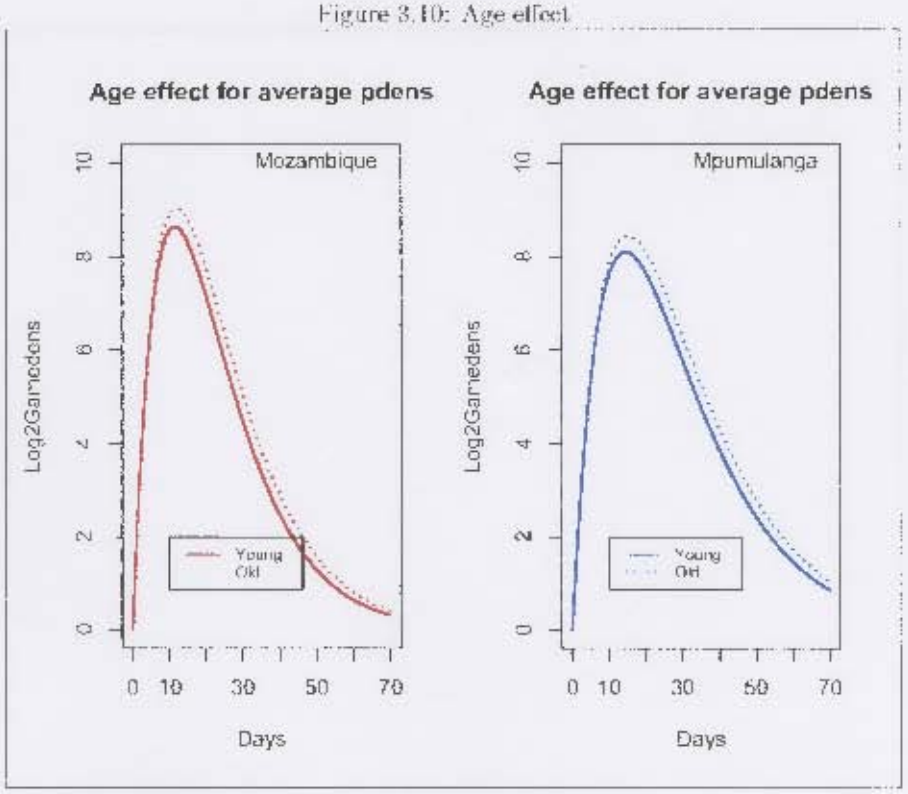
Figure 3.9: Site effect



- Age: A patient that is one year older then another patient with all else being equal, is expected to have a value for R that is approximately 0.0002 units higher (or 0.002 units for every 10 years). Plot 3.10 below depicts the age effect by site. This shows that older people are expected to have a slightly higher peak that occurs slightly later.

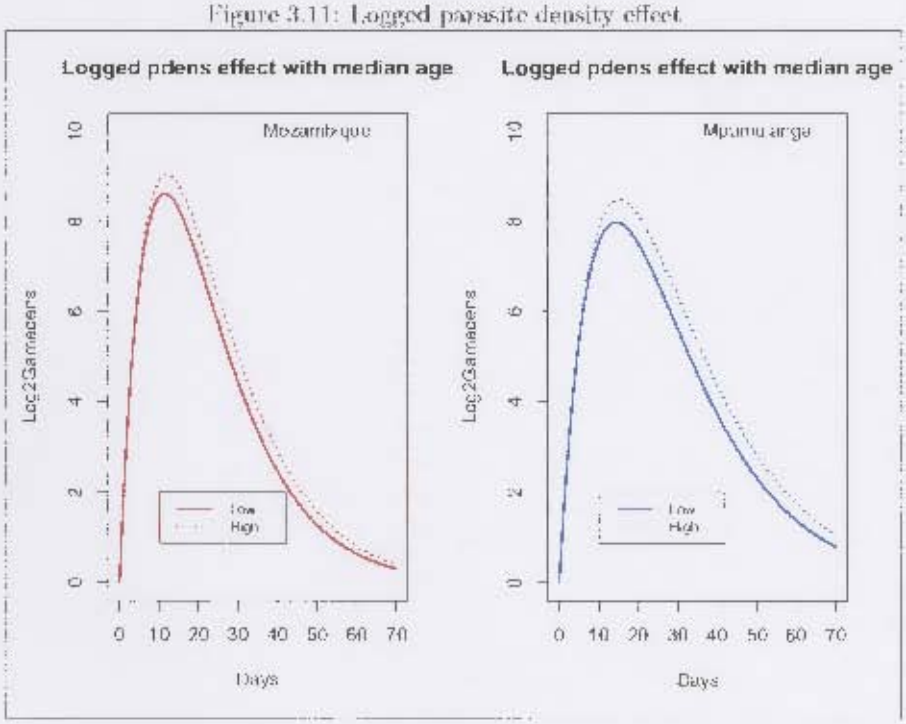
  Note that the first and third quartiles within each site were used here in the definition of 'Young' and 'Old' subjects and since they are not the same, the two sites are not directly comparable. Specifically the 1st quartile of age in Mozambique is lower than Mpumulanga (5 vs 11). The third quartile is virtually the same (29 vs 28.5).

Figure 3.10: Age effect



- Logged asexual parasite density at baseline: A patient with a pre-treatment logged parasite density that is one unit higher than another patient with all else being equal, is expected to have a value for R that is 0.006 units higher. Note that a one unit increase here is a ten fold increase on the original scale. Plot 3.11 below depicts the logged parasite density effect by site and shows that the effect is similar to the age effect above. Note that the overall first and third quartiles were used here in the definition of 'Low' and 'High' logged parasite densities.

Figure 3.11: Logged parasite density effect

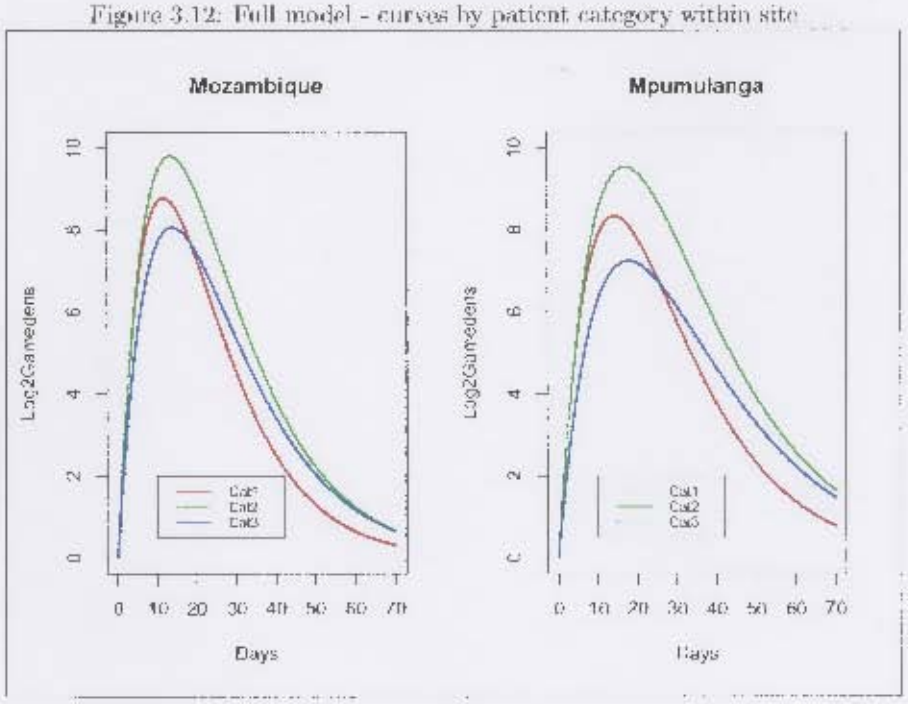


- Patient category: A patient who falls in the second category is expected to have a value for C that is about 0.06 units lower and a value for R that is 0.01 units higher when compared to a patient in category one. The different value for C between categories two and one was not found to be significantly different.

  On the other hand a patient who falls in the 3rd category is expected to have a value for C that is about 0.49 units lower and a value for R that is 0.013 units higher when compared to a patient in category one. Both these effects were significantly different to patients in the first category.

  Plot 3.12 below depicts the curves for each site seperately and by patient category.

Figure 3.12: Full model - curves by patient category within site



The model predicts that patients in category two, on average, have a higher and later peak than category one but with a very similar rate of increase and decrease. On the other hand, the model predicts that patients in the 3rd category would have a lower peak that occurs slightly later and then clears more slowly i.e. these patients would carry gametocytes for longer. This makes sense when one considers that patients in the 3rd category have a second wave of gametocytes or two seperate peaks, and that the function used in the model copes with this by smoothing out these peaks thereby resulting in a lower average peak and a slower clearance rate.

So to conclude, patients from Mozambique tend to have higher gametocyte densities but a faster rate of clearance compared to patients from Mpumulanga, while older people and people with high day 0 asexual parasite counts tend to have higher gametocyte densities.

Referring to figure 1.6 in section 1, the objective of the modeling was to produce gametocyte density-time profiles so that firstly gametocyte densities and secondly gametocyte clearance rates at a particular time for particular patient types could be estimated. The fact that the ratio between the responses at successive time points is not equal in this function also means that the clearance

rate will change with time. The non-constant clearance rate, together with the fact that the chosen function does not have an underlying biological interpretation for the parameters, is a weakness of the current model. However, in a larger model of malaria transmission, gametocyte clearance rates could be modelled as a time-varying covariate and hence these curves could still be utilised.

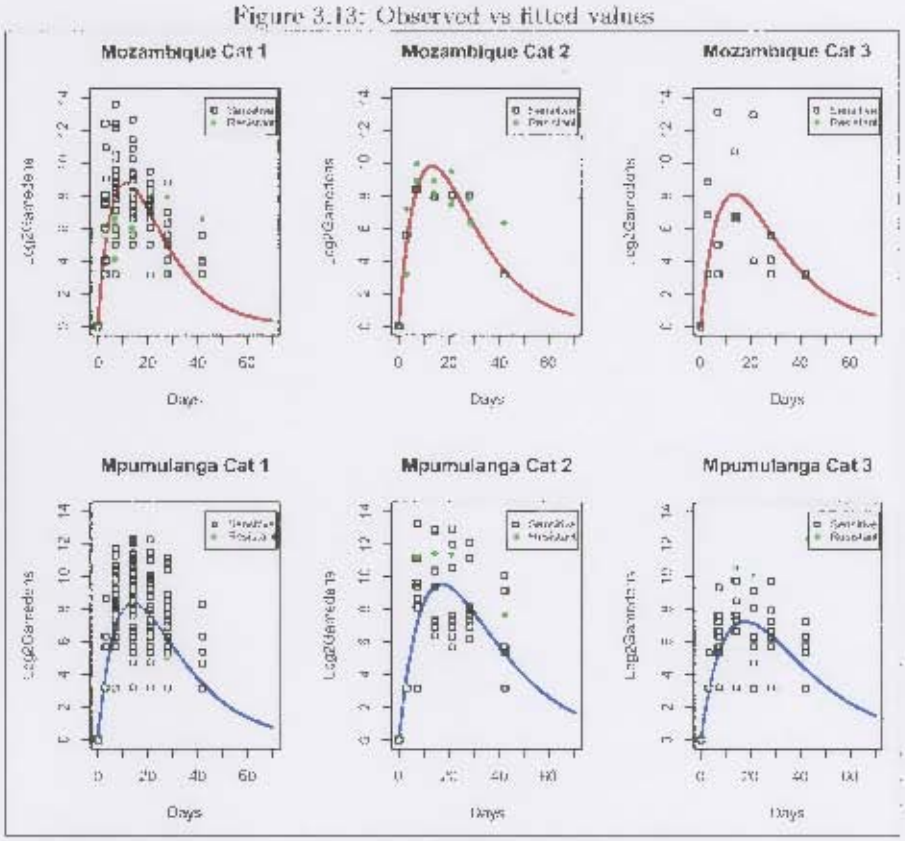
Model fit

Various plots were generated in order to investigate the fit of the final model (including the patient category variable). The code for this section can be found in Appendix D.2.3. Plots from the final model showing the individual fits along with the population averages are shown in plot C.1 found in appendix C. The fixed curves represent population curves based on the fixed effects whereas the curves displayed with broken lines (labelled subjectno) represent curves for individual patients that are obtained by adding random effects to the fixed effects. Plot C.1 shows that the population curves in general approximate the observed data fairly well and how adding random effects improves the fit for the different individuals. The chosen function appears capable of taking on a wide variety of shapes.

Plot 3.13 below depicts the observed values against population averages, as estimated by the final model, for each site and patient category. Overall the fitted lines appear to pass through the middle of the observed points with the exception of the first category where the line seems to be a bit low for the latter part of the infection. The points have also been labelled according to the mutation variable in order to investigate any systematic occurence of the resistant mutations, however there is nothing apparent from the plot.

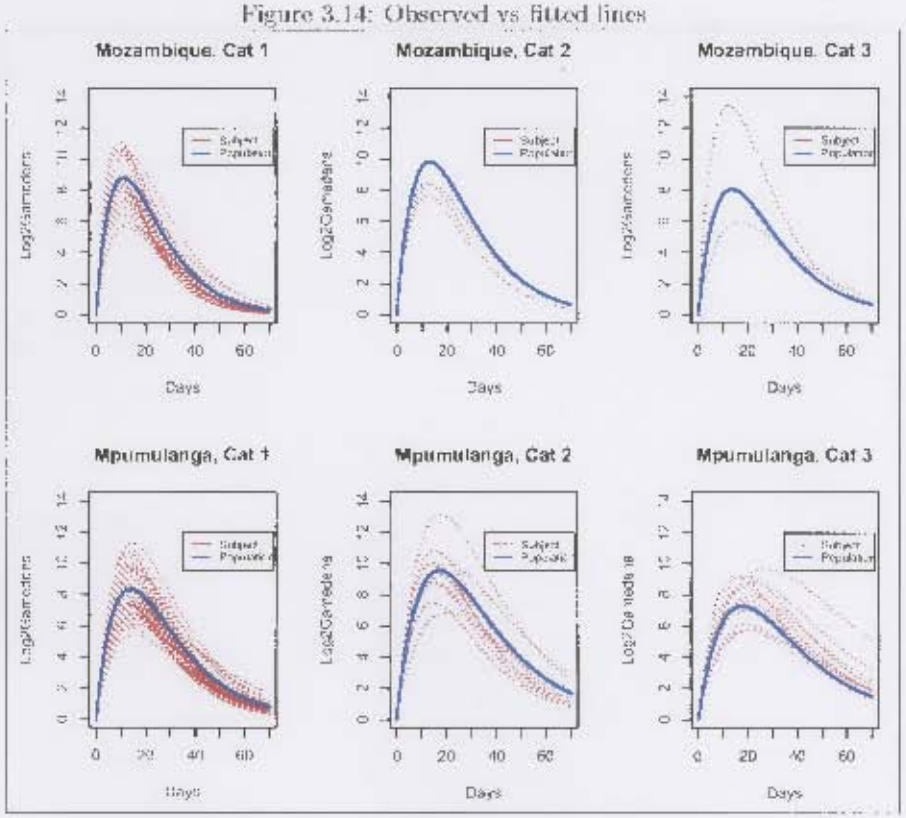
Note that the data only go up to day 42 and hence the shapes of the population curves beyond day 42 are effectively extrapolation in both graphs. One could present the greater uncertainty associated with estimates beyond day 42 by including confidence bands but this was not done here.

Figure 3.13: Observed vs fitted values



The fitted lines for each subject were plotted against the population average by site and patient category in the next plot (plot 3.14). Plot 3.14 confirms the large amount of individual variability in the gametocyte density profiles. It is also evident that, with the exception of category two for Mozambique, the fitted line is an average of the subject lines. The exception can be explained by the low number of patients in that group thereby leading to a shrinkage towards the overall mean.

Figure 3.14: Observed vs fitted lines



Chapter 4

Discussion and Conclusions

Host infectivity is a critical component in the transmission of malaria and the spread of resistance. While various factors are thought to influence host infectivity, the best indicator appears to be gametocyte carriage. Many previous or existing models for the transmission of malaria have thus implemented infectivity from gametocyte density.

Yet despite this, my review of the literature has not found a single case where gametocytes themselves have been modelled directly. Rather, it appears that focus is usually on the more frequently measured asexual parasite densities from which gametocyte carriage is estimated by either making a simple assumption (such as using a constant gametocyte switching rate), or by using a more advanced model i.e. using lagged values of asexual parasite densities. However, the literature suggests that the relationship between asexual and sexual densities is far from conclusive and this raises doubts around estimating gametocytaemia based only on asexual densities.

The huge obstacle that drug resistance poses for the control of malaria together with the role that gametocytes play in the spread of such resistance makes it critical to examine gametocyte carriage more closely. The model developed in this thesis aims to make a contribution by capturing host infectivity in a more realistic and accurate way by modelling gametocytes directly. It could then be used in a larger model of malaria transmission and the spread of resistance.

Unfortunately, the study design from which this data was obtained used measurement points that were based around the expected clinical and asexual parasitological response to malaria treatment. This is embodied in the fact that patients were seen fairly intensively during the first week of their infection and then less intensively later on i.e. a gap of 14 days between the last two time points. Since the distribution of gametocytes always lags that of asexual parasites the current design is not optimal for measuring gametocytes and hence the current data cannot support complicated functions. The problem is exac-

erbated by the withdrawal of patients who fail treatment so that they can be given rescue treatment. In addition, duration of symptoms prior to treatment was not recorded.

Despite these limitations, the model developed with the modified critical exponential function is promising as it seems able to encompass a wide range of variability in the curves, and appears to be stable across the different datasets. Essentially this means that the chosen function is robust to the inclusion/exclusion of different patient category types as well as the amount of information available on the response for the different patients. This function was preferred to the one compartment model, primarily due to its performance with Data2 when the zeros were replaced with 8's and resulted in more realistic right-hand tails that correspond to faster clearance of gametocytes. The disadvantage of choosing this function is the lack of clear and distinct interpretations for the parameters as opposed to a more mechanistic function like the one compartment model.

A richer dataset with more optimally chosen measurement intervals as well as follow up of gametocyte carriage in treatment failures, would allow more complex functions to be examined and could potentially lead to an improvement on the function used in this analysis, particularly with regards to finding parameters that have more meaningfull biological/clinical interpretations.

Section 3.4.1 suggests that the wave-like double fourier warrants further investigation with a richer dataset. While this function could be useful there are questions around the importance of capturing a wave-like pattern in the data. The process of sequestration may lead to the observed gametocyte densities not really being reflective of the true total gametocyte burden (seeing as there are gametocytes sticking to the endothelium that cannot be detected). Furthermore, since the recurrent wave that is observed in some gametocyte distributions appears to be related to the synchronicity of the infection which varies between patients, the wave-like structure would not be observed for all patients unless they all carried synchronous infections. It could therefore make more sense to estimate curves that 'smooth' through these waves much like what was observed with this data, and hence attempt to estimate the true level of gametocytaemia.

The current model that uses the modified critical exponential function would not manage to model gametocyte density-time profiles for people that are expected to carry gametocytes for long periods of time where the gametocytes continue to recur in waves. It is obvious that one would need a wave-like function for this sort of scenario.

Various covariates were useful at explaining some of the random variation in parameters between individuals. The site effect could reflect a degree of immunity as malaria transmission in Mozambique is more intense than in Mpumulanga. Immunity could trigger an early release of gametocytes and consequently explain the faster increase to an earlier peak as well as a faster clearance rate.

The effect of age is less easy to explain but it is feasible that adults are more vigilant with regards to seeking treatment for their children compared to for themselves. This would result in children being treated earlier than adults and could thus explain the slightly higher densities for adults. As expected, higher asexual parasite densities at baseline are related to higher gametocyte densities.

The patient category variable was clearly important and could be related to the synchronicity of infection. The fact that some association between mutations and patient category was found in the data exploration is not surprising, since a resistant infection is likely to recrudesce and the fact that the duration of infection would be longer allows the gametocytes to start developing in synch with each other. It was interesting that there was very little difference in the interpretation of the other covariate parameters when excluding patient category. This suggests that the patient categories can be viewed as a proxy for characteristics that were not measured yet were also independent from the available covariate information.

The literature suggests that resistant infections increase gametocytaemia and hence it was surprising that the variable measuring mutations did not enter the model. I investigated this using a redefined mutation variable (with mixed infections kept together) but still could not detect any significant effect. It is therefore conjectured that this is related to power and the fact that there were very few patients in the data who had resistant infections. It is possible that with more power, mutations could even explain the patient category variable and hence be used in its place.

Directly related to the lack of power is the fact that most treatment failures are rescued and removed from the study. This makes it very difficult to accurately estimate these gametocyte density-time profiles. Either a questionable assumption (such as equating their curves with treatment successes) has to be made, or better data that tracks their gametocytes after being withdrawn from the study needs to be obtained.

Lastly, this research is limited to carriers of gametocytes. In order to model the dynamics of malaria transmission and the spread of resistance in a population, one would need to also incorporate a component that captures the determinants of gametocyte prevalence i.e. what causes some people to present with gametocytes while others do not.

Bibliography

Aron, J. L. (1988). Mathematical modeling of immunity to malaria. *Mathematical Biosciences*, 90:385–396.

Aron, J. L. and May, R. M. (1982). The poulation dynamics of malaria. In Anderson, R. M., editor, *Population Dynamics of Infectious Diseases*, Population and Community Biology, chapter 5, pages 139–179. Chapman and Hall, London.

Babiker, H. A., Satti, G., Ferguson, H., Bayoumi, R., and Walliker, D. (2005). Drug resistant plasmodium falciparum in an area of seasonal transmission. *Acta Tropica*, 94(3):260–268.

Bailey, N. (1975). *The mathematical theory of infectious diseases and its applications*. Charles Griffin and Company Limited, London.

Baird, J. K. (1998). Age-dependent characteristics of protection v. susceptibility to plasmodium falciparum. *Annals of Tropical Medicine and Parasitology*, 92(4):367–390.

Baird, J. K., Jones, T. R., Danudirgo, E. W., Annis, B. A., Bangs, M. J., Basri, H., Purnomo, and Masbar, S. (1991). Age-dependent acquired protection against plasmodium falciparum in people having two years exposure to hyperendemic malaria. *American Journal of Tropical Medicine and Hygiene*, 45(1):65–76.

Barnes, K. and White, N. J. (2005). Population biology and antimalarial resistance: The transmission of antimalarial drug resistance in plasmodium falciparum. *Acta Tropica*, 94(3):230–240.

Becker, N. G. (1989). *Analysis of Infectious Disease Data*. Chapman and Hall.

Bredenkamp, B. L., Sharp, B. L., Mthembu, S. D., Durrheim, D. N., and Barnes, K. I. (2001). Failure of sulphadoxine-pyrimethamine in treating plasmodium falciparum malaria in kwazulu-natal. *SAMJ*, 91(11):970–972.

Craig, M. H., Kleinschmidt, I., Sueur, D. L., and Sharp, B. L. (2004). Exploring 30 years of malaria case data in kwazulu-natal, south africa: Part ii. the impact of non-climatic factors. *Tropical medicine and International Health*, 9(12):1258–1266.

Davidian, M. and Giltinan, D. (1998). *Nonlinear models for repeated measurement data.* Chapman & Hall.

Diebner, H. H., Eichner, M., Molineaux, L., Collins, W. E., Jeffery, G. M., and Dietz, K. (2000). Modelling the transition of asexual blood stages of plasmodium falciparum to gametocytes. *Journal of Theoretical Biology*, 202:113–127.

Dietz, K. (1988). Mathematical models for transmission and control of malaria. pages 1091–1133. Churchill Livingstone, Edinburgh.

Dietz, K., Molineaux, L., and Thomas, A. (1974). A malaria model tested in the african savannah. *Bulletin of the World Health Organisation*, 50:347–357.

Dietz, K., Raddatz, G., and Molineaux, L. (2006). Mathematical model of the first wave of plasmodium falciparum asexual parasitemia in non-immune and vaccinated individuals. *American Journal of Tropical Medicine and Hygiene*, 75(2 Supplement):46–55.

Drakeley, C. J., Jawara, M., Targett, G. A., Walraven, G., Obisike, U., Coleman, R., Pinder, M., and Sutherland, C. J. (2004). Addition of artesunate to chloroquine for treatment of plasmodium falciparum malaria in gambian children causes a significant but short-lived reduction in infectiousness for mosquitoes. *Tropical medicine and International Health*, 9(1):53–61.

Drakeley, C. J., Secka, I., Correa, S., Greenwood, B. M., and Targett, G. A. T. (1999). Host haematological factors influencing the transmission of plasmodium falciparum gametocytes to anopheles gambiae s.s. mosquitoes. *Tropical medicine and International Health*, 4(2):131–138.

Draper, C. C. (1953). Observations on the infectiousness of gametocytes in hyper-endemic malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 47(2):160–165.

Fine, P. E. (1975). Ross's *a priori* pathometry - a perspective. *Journal of the Royal Society of Medicine*, 68(9):547–551.

Garrett-Jones, C. (1964). The human blood index of malaria vectors in relation to epidemiological assessment. *Bulletin of the World Health Organisation*, 30:241–261.

Graves, P. M., Carter, R., Burkot, T. R., Quakyi, I. A., and Kumar, N. (1988). Antibodies to plasmodium falciparum gamete surface antigens in papua new guinea sera. *Parasite Immunology*, 10:209–218.

Greenwood, B. M. (1997). The epidemiology of malaria. *Annals of Tropical Medicine & Parasitology*, 91(7):763–769.

Gu, W., Killeen, G. F., Mbogo, C. M., Regens, J. L., Githure, J. I., and Beier, J. C. (2003a). An individual-based model of plasmodium falciparum malaria transmission on the coast of kenya. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 97:43–50.

Gu, W., Mbogo, C., Githure, J., Regens, J., Killeen, G., Swalm, C., Yan, G., and Beier, J. (2003b). Low recovery rates stabilise malaria endemicity in areas of low transmission in coastal kenya. *Acta Tropica*, 86:71–81.

Gupta, S., Snow, R. W., Donnelly, C. A., Marsh, K., and Newbold, C. (1999). Immunity to non-cerebral severe malaria is aquired after one or two infections. *Nature Medicine*, 5(3):340–343.

Haji, H., Smith, T., Charlwood, J. D., and Meuwissen, J. H. (1996). Absence of relationships between selected human factors and natural infectivity of plasmodium falciparum tomosquitoes in an area of high infection. *Parasitology*, 113:425–431.

Heesterbeek, J. A. P. (2002). A brief history of r_0 and a recipe for its calculation. *Acta Biotheoretica*, 50:189–204.

Hogh, B., Gamage-Mendis, A., Butcher, G. A., Thompson, R., Begtrup, K., Mendis, C., Enosse, S. M., Dgedge, M., Barreto, J., Eling, W., and Sinden, R. (1998). The differing impact of chloroquine and pyrimethamine/sulfadoxine upon the infectivity of malaria species to the mosquito vector. *American Journal of Tropical Medicine and Hygiene*, 58(2):176–182.

Jeffery, G. M. and Eyles, D. E. (1955). Infectivity to mosquitoes of plasmodium falciparum as related to gametocyte density and duration of infection. *American Journal of Tropical Medicine and Hygiene*, 4:781–789.

Killeen, G., Ross, A., and Smith, T. (2006). Infectiousness of malaria-endemic human populations to vectors. *American Journal of Tropical Medicine and Hygiene*, 75(2 Supplement):38–45.

Kleinschmidt, I. (2001). *Spatial statistical analysis, modelling and mapping of malaria in Africa*. PhD thesis, University of Basel.

Koella, J. C. (1991). On the use of mathematical models of malaria transmission. *Acta Tropica*, 49:1–25. Review.

Korenromp, E. L., Williams, B. G., Gouws, E., Dye, C., and Snow, R. (2003). Measurement of trends in childhood malaria mortality in africa: an assessment of progress toward targets based on verbal autopsy. *The Lancet infectious diseases*, 3:349–358.

Luxemburger, C., Ricci, F., Nosten, F., Raimond, D., Bathet, S., and White, N. J. (1997). The epidemiology of severe malaria in an area of low transmission in thailand. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 91:256–262.

MacDonald, G. (1950). The analysis of infection rates in diseases in which superinfection occurs. *Bureau of Hygiene and Tropical Diseases*, 47(10):907–937.

MacDonald, G. (1957). *The Epidemiology and Control of Malaria.* Oxford University Press, London.

Maire, N., Smith, T., Ross, A., Owusu-Agyei, S., Dietz, K., and Molineaux, L. (2006). A model for natural immunity to asexual blood stages of plasmodium falciparum malaria in endemic areas. *American Journal of Tropical Medicine and Hygiene*, 2 Supplement:19–31.

Mayxay, M., Chotivanich, K., Pukrittayakamee, S., Newton, P., Looareesuwan, S., and White, N. J. (2001). Contribution of humoral immunity to the therapeutic response in falciparum malaria. *American Journal of Tropical Medicine and Hygiene*, 65(6):918–923.

McKenzie, F. E. and Bossert, W. H. (1998). The optimal production of gametocytes by plasmodium falciparum. *Journal of Theoretical Biology*, 193:419–428.

Mendis, K. N., Munesinghe, Y. D., Silva, Y. N. Y. D., Keragalla, I., and Carter, R. (1987). Malaria transmission-blocking immunity induced by natural infections of plasmodium vivax in humans. *Infection and Immunity*, 55(2):369–372.

Molineaux, L. (1985). The pros and cons of modelling malaria transmission. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 79(6):743–747.

Muheki, C., McIntyre, D., and Barnes, K. (2004). Artemisinin-based combination therapy reduces expenditure on malaria treatment in kwazulu natal, south africa. *Tropical medicine and International Health*, 9(9):959–966.

Mulder, B., Tchuinkam, T., Dechering, K., Verhave, J. P., Carnevale, P., Meuwissen, J., and Robert, V. (1994). Malaria transmission-blocking activity in experimental infections of anopheles gambiae from naturally infected plasmodium falciparum gametocyte carriers. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 88:121–125.

Nosten, F., van Vugt, M., Luxemburger, C., Thway, K. L., Brockman, A., McGready, R., ter Kuile, F., Looareesuwan, S., and White, N. J. (2000). Effects of artesunate-mefloquine combination on incidence of plasmodium falciparum malaria and mefloquine resistance in western thailand: a prospective study. *The Lancet*, 356:297–302.

Pinheiro, J. and Bates, D. (2002). *Mixed-effects Models in S and S-PLUS.* Statistics and Computing. Springer-Verlag, New York.

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2006). nlme: Linear and nonlinear mixed effects models. R package version 3.1-68.1.

Pongtavornpinyo, W. (2006). *Mathematical modelling of antimalarial drug resistance.* PhD thesis, School of Tropical Medicine, University of Liverpool.

Price, R., Nosten, F., Simpson, J., Luxemburger, C., Phaipun, L., ter Kuile, F., Vugt, M. V., Chongsuphajaisiddhi, T., and White, N. (1999). Risk factors for gametocyte carriage in uncomplicated falciparum malaria. *American Journal of Tropical Medicine and Hygiene*, 60(6):1019–1023.

Price, R. N., Nosten, F., Luxemburger, C., ter Kuil, F. O., Paiphun, L., Chongsuphajaisiddhi, T., and White, N. J. (1996). Effects of artemisinin derivatives on malaria transmissibility. *Lancet*, 347:1654–1658.

R Development Core Team (2005). R: A language and environment for statistical computing. http://www.R-project.org.

Ratkowsky, D. (1990). *Handbook of nonlinear regression models.* M. Dekker, New York.

Robert, V., Awono-Ambene, H. P., Hesran, J. Y. L., and Trape, J. F. (2000). Gametocytemia and infectivity to mosquitoes of patients with uncomplicated *Plasmodium falciparum* malaria attacks treated with chloroquine or sulfadoxine plus pyrimethamine. *American Journal of Tropical Medicine and Hygiene*, 62(2):210–216.

Rogier, C., Ly, A. B., Tall, A., Cisse, B., and Trape, J. F. (1999). Plasmodium falciparum clinical malaria in dielmo, a holoendemic area in senegal: No influence of acquired immunity on initial symptomatology and severity of malaria attacks. *American Journal of Tropical Medicine and Hygiene*, 60(3):410–420.

Roll Back Malaria (2006). The roll back malaria partnership. http://www.rollbackmalaria.org/ Last accessed January 2007.

Roper, C., Pearce, R., Bredenkamp, B., Gumede, J., Drakeley, C., Mosha, F., Chandramohan, D., and Sharp, B. (2003). Antifolate antimalarial resistance in southeast africa: a population-based analysis. *The Lancet*, 361:1174–81.

Ross, A., Killeen, G., and Smith, T. (2006). Relationships between host infectivity to mosquitoes and asexual parasite density in *Plasmodium falciparum*. *American Journal of Tropical Medicine and Hygiene*, 75(2 Supplement):32–37.

Ross, R. (1911). *The Prevention of Malaria.* Murray, London.

Sachs, J. and Malaney, P. (2002). The economic and social burden of malaria. *Nature*, 415(6872):680–685.

Simpson, J. A., Aarons, L., Collins, W. E., Jeffery, G. M., and White, N. J. (2002). Population dynamics of untreated plasmodium falciparum malaria within the adult human host during the expansion phase of the infection. *Parasitology*, 124:247–263.

Smith, D. and McKenzie, F. (2004). Statics and dynamics of malaria infection in anopheles mosquitoes. *Malaria Journal*, 13(3). Review.

Smith, T., Killeen, G. F., Maire, N., Ross, A., Molineaux, L., Tediosi, F., Hutton, G., Utzinger, J., Dietz, K., and Tanner, M. (2006). Overview. *American Journal of Tropical Medicine and Hygiene*, 75(2 Supplement):1–10.

Snow, R. W., Craig, M., Deichmann, U., and Marsh, K. (1999a). Estimating mortality, morbidity and disability due to malaria among africa's non-pregnant population. *Bulletin of the World Health Organisation*, 77(8):624–640.

Snow, R. W., Craig, M. H., Deichmann, U., and le Seur, D. (1999b). A preliminary continental risk map for malaria mortality among african children. *Parasitology Today*, 15(3):99–104.

Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y., and Hay, S. (2005). The global distribution of clinical episodes of plasmodium falciparum malaria. *Nature*, 434:214–217.

Snow, R. W. and Marsh, K. (1998). New insights into the epidemiology of malaria relevant for disease control. *British Medical Bulletin*, 54(2):293–309.

Snow, R. W., Nahlen, B., Palmer, A., Donnelly, C. A., Gupta, S., and Marsh, K. (1998). Risk of severe malaria among african infants: Direct evidence of clinical protection during early infancy. *Journal of Infectious Diseases*, 177:819–822.

Snow, R. W., Trape, J. F., and Marsh, K. (2001). The past, present and future of childhood malaria mortlity in africa. *TRENDS in Parasitology*, 17(12):593–597.

StataCorp, College Station, Texas (2005). Stata statistical software: Release 8. http://www.stata.com.

Targett, G., Drakeley, C., Jawara, M., von Seidlein, L., Coleman, R., Deen, J., Pinder, M., Doherty, T., Sutherland, C., Walraven, G., and Milligan, P. (2001). Artesunate reduces but does not prevent posttreatment transmission of plasmodium falciparum to anopheles gambiae. *Journal of Infectious Diseases*, 183:1254–1259.

Tchuinkam, T., Mulder, B., Dechering, K., Stoffels, H., Verhave, J. P., Cot, M., Carnevale, P., Meuwissen, J. H., and Robert, V. (1993). Experimental infections of anopheles gambiae with plasmodium falciparum of naturally infected gametocyte carriers in cameroon: factors influencing the infectivity to mosquitoes. *Tropical Medicine and Parasitology*, 44(4):271–276.

Thompson, J. R. (1989). *Empirical Model Building.* Wiley, New York.

Thompson, J. R. (2000). *Simulation: a modeller's approach.* Wiley, New York.

Trape, J. F. (2001). The public health impact of chloroquine resistance in africa. *American Journal of Tropical Medicine and Hygiene*, 64(Supplement):12–17.

Verhoeff, F. H., Brabin, B. J., Hart, C. A., Chimsuku, L., Kazembe, P., and Broadhead, R. L. (1999). Increased prevalence of malaria in hiv-infected pregnant women and its implications for malaria control. *Tropical medicine and International Health*, 4(1):5–12.

White, N. J. (2002). The assessment of antimalarial drug efficacy. *TRENDS in Parasitology*, 18(10):458–464.

White, N. J. (2004). Antimalarial drug resistance. *The Journal of Clinical Investigation*, 113(8):1084–1092.

White, N. J., Chapman, D., and Watt, G. (1992). The effects of multiplication and synchronicity on the vascular distribution of parasites in falciparum malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 86(6):590–597.

Whitworth, J., Morgan, D., Quigley, M., Smith, A., Mayanja, B., Eotu, H., Omoding, N., Okongo, M., Malamba, S., and Ojwiya, A. (2000). Effect of hiv-1 and increasing immunosuppression on malaria parasitaemia and clinical episodes in adults in rural uganda: a cohort study. *The Lancet*, 356:1051–56.

WHO Press (2006). Who guidelines for the treatment of malaria. Website. http: www.who.int.

Appendix A

Correlation Matrices

Table A.1: Data2(56), Correlations from fourier model

A	A	B	E
B	0.888		
E	0.472	0.013	
W	-0.544	-0.097	-0.996

Table A.2: Data2(103), Correlations from fourier model

A	A	B
B	0.961	
E	0.423	0.158

Table A.3: Data1(103), Correlations from the double fourier model

A	A	B	E	W	G
B	0.664				
E	0.383	-0.412			
W	-0.503	0.247	-0.81		
G	0.795	0.942	-0.197	0.074	
P	0.343	-0.446	0.906	-0.96	-0.282

Appendix A

Correlation Matrices

Table A.1: Data2(56), Correlations from fourier model

A	A	B	E
B	0.888		
E	0.472	0.013	
W	-0.544	-0.097	-0.996

Table A.2: Data2(103), Correlations from fourier model

A	A	B
B	0.961	
E	0.423	0.158

Table A.3: Data1(103), Correlations from the double fourier model

A	A	B	E	W	G
B	0.664				
E	0.383	-0.412			
W	-0.503	0.247	-0.81		
G	0.795	0.942	-0.197	0.074	
P	0.343	-0.446	0.906	-0.96	-0.282

Table A.4: Data2(56), Correlations from the double fourier model

A	A	B	E	W	G
B	0.612				
E	0.536	-0.337			
W	-0.514	0.288	-0.937		
G	0.560	0.992	-0.397	0.375	
P	0.436	-0.437	0.991	-0.953	-0.501

Table A.5: Data2(103), Correlations from the double fourier model

A	A	B	E	W	G
B	0.657				
E	0.452	-0.358			
W	-0.347	0.378	-0.885		
G	0.648	0.863	-0.297	0.408	
P	0.340	-0.419	0.948	-0.977	-0.456

Appendix B

Model Building Plots

Note that the variable called mut_3 refers to the mutation variable. The plots below show a 3rd category for mutations as there was no mutation data for the full dataset and hence this category is for the patients with missing data on mutations. Also note that the labels for the treatment outcome variable *fail* are presented below as "Resistant" and "Sensitive" which correspond to "Treatment failure" and Treatment success" respectively.

Figure B.1: Data1, C - No covariates

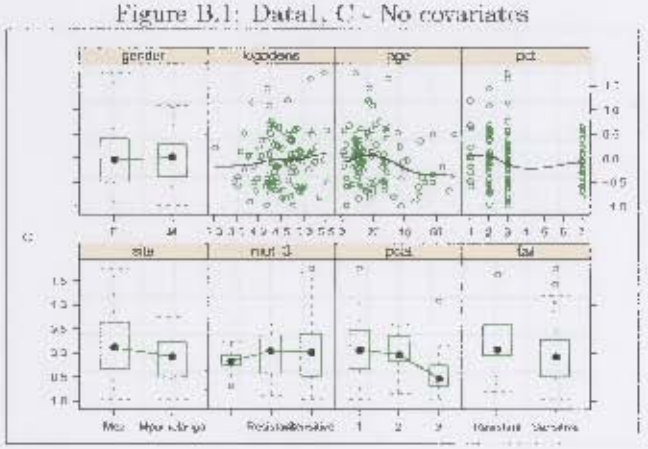


Figure B.2: Data1, R - No covariates

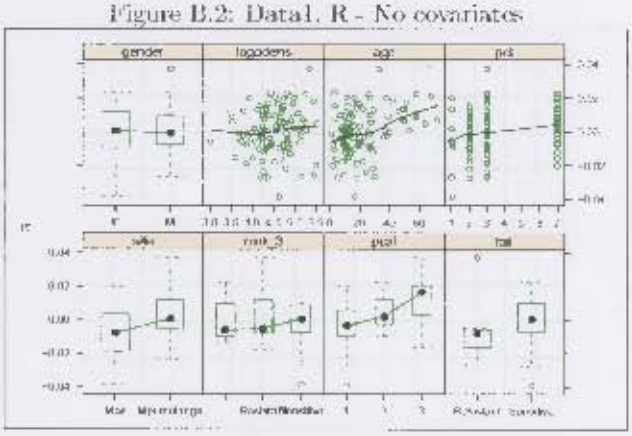


Figure B.3: Data1, C - Age

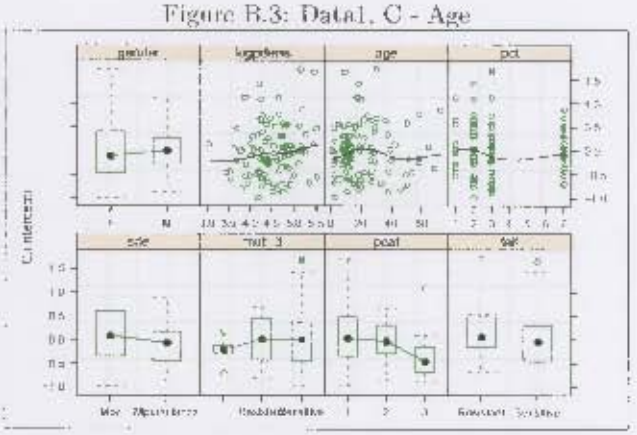


Figure B.4: Data1, R - Age

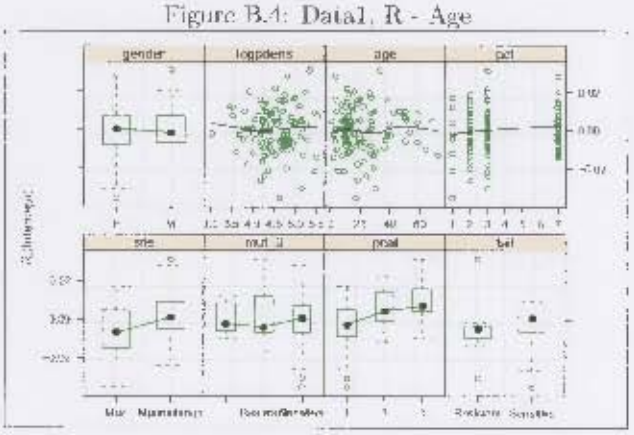


Figure B.5: Data1, C - Age, Site

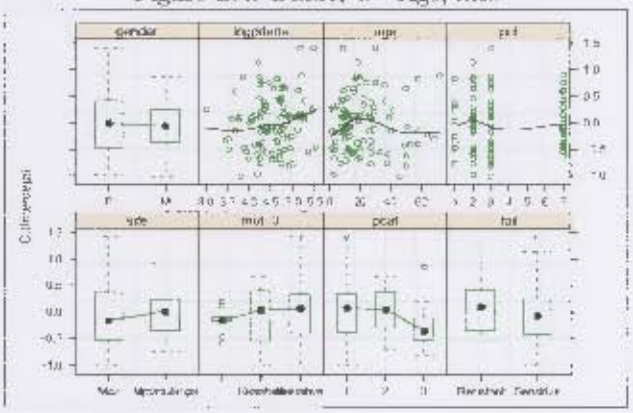


Figure B.6: Data1, R - Age, Site

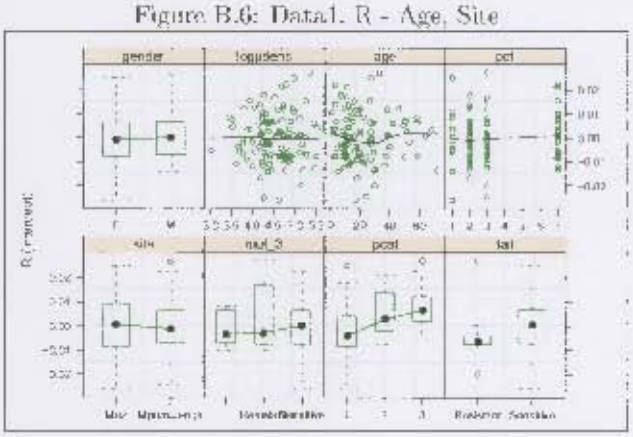


Figure B.7: Data1, C - Age, Site, Logpdens

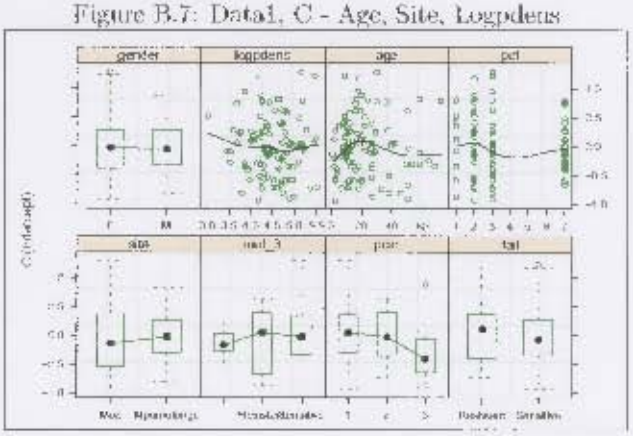


Figure B.8: Data1, R - Age, Site, Logpdens

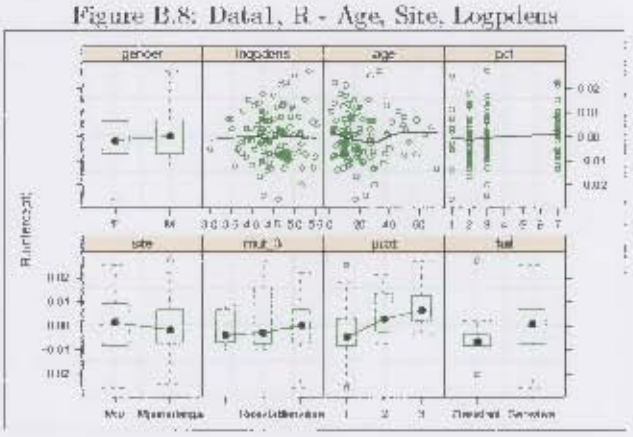


Figure B.9: Data1, C - Age, Site, Logpdens,pcat

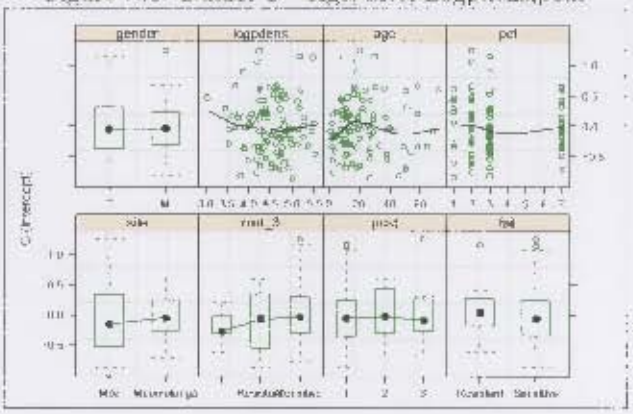


Figure B.10: Data1, R - Age, Site, Logpdens,pcat

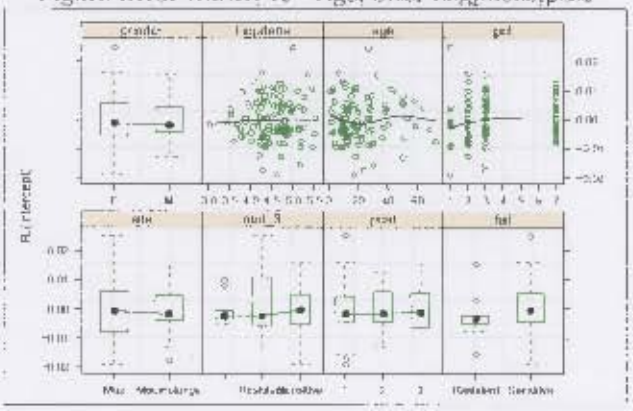


Figure B.11: Data1, Diagnostics - qnorm

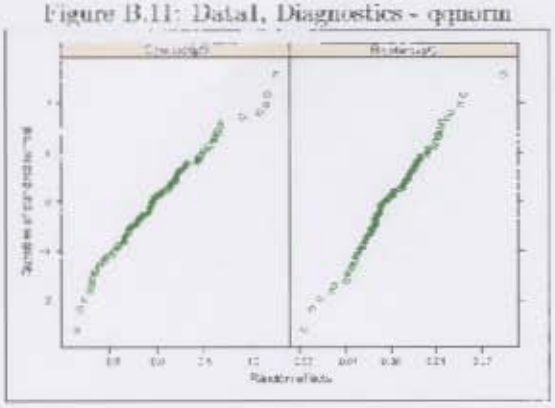


Figure B.12: Data1, Diagnostics - residuals

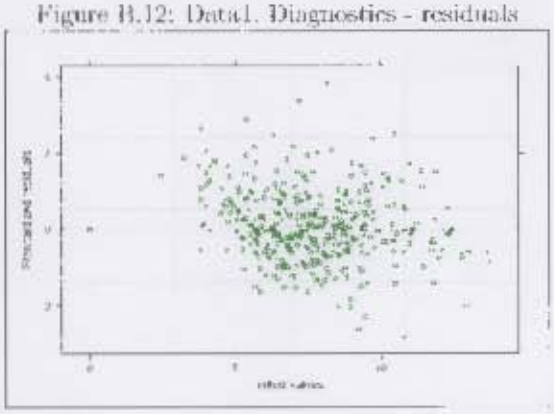


Figure B.13: Data1, Diagnostics - fitted vs observed

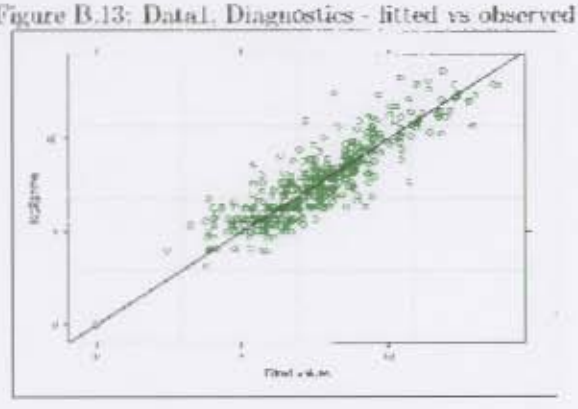


Figure B.14: Data2, C - No covariates

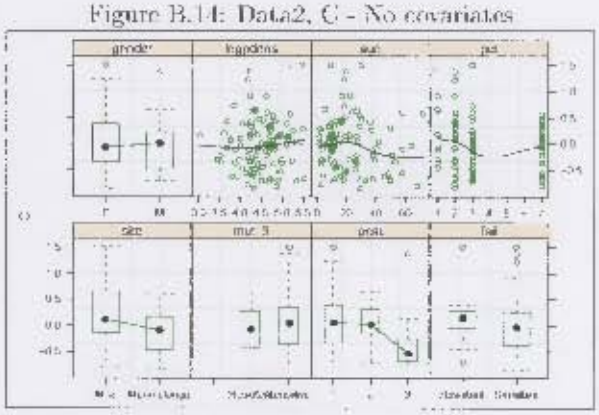


Figure B.15: Data2, R - No covariates

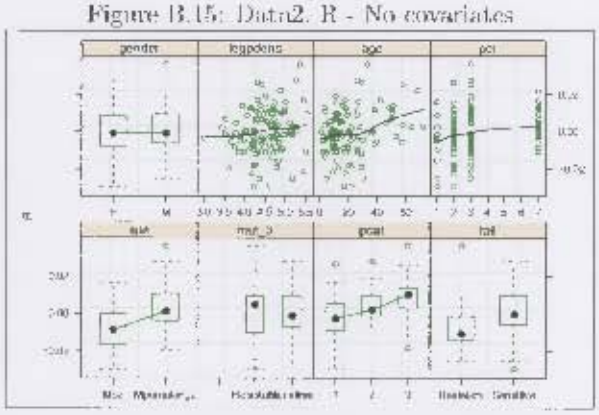


Figure B.16: Data2, C - Age (on R)

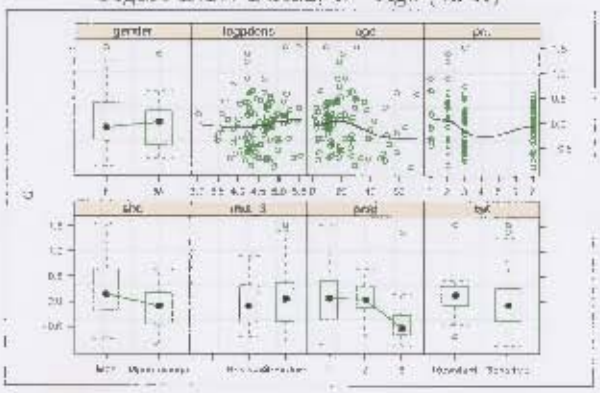


Figure B.17: Data2, R - Age

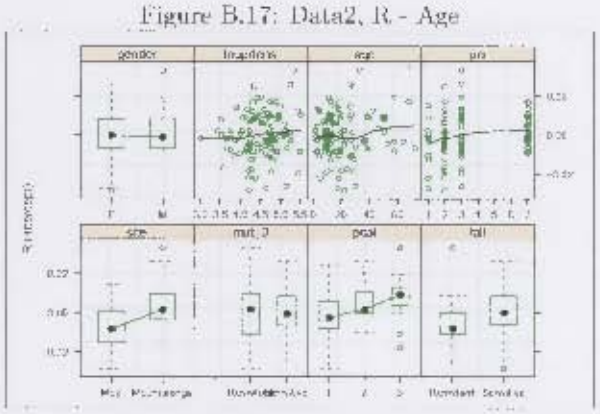


Figure B.18: Data2, C - Age (on R), Site

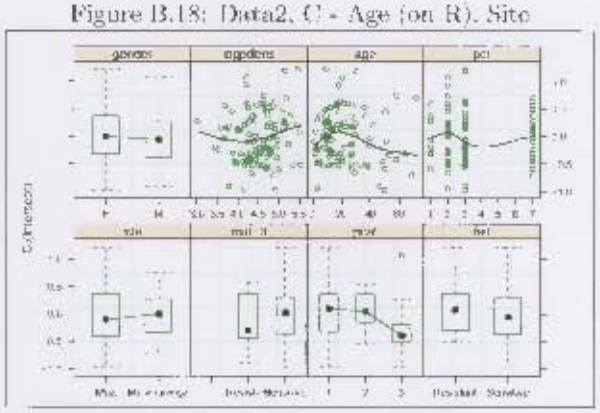


Figure B.19: Data2, R - Age, Site

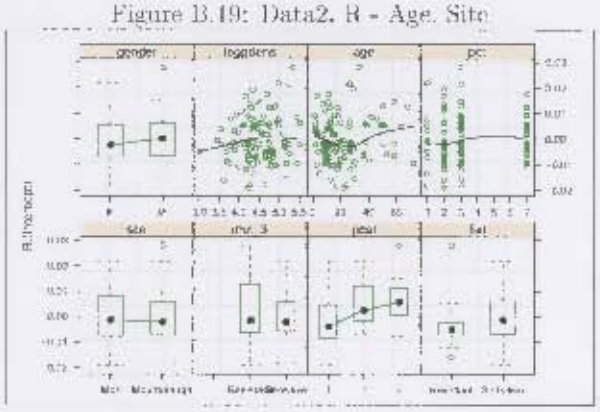


Figure B.20: Data2, C - Age (on R), Site, Logpdens

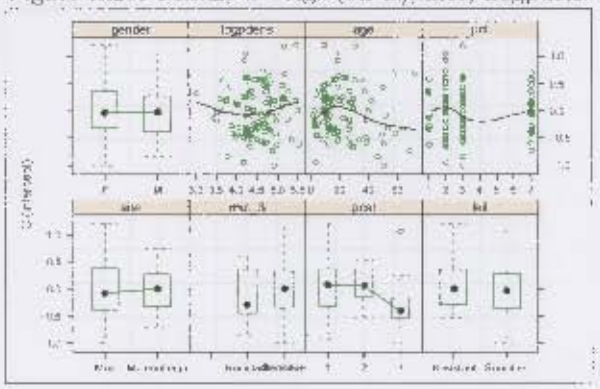


Figure B.21: Data2, R - Age, Site, Logpdens

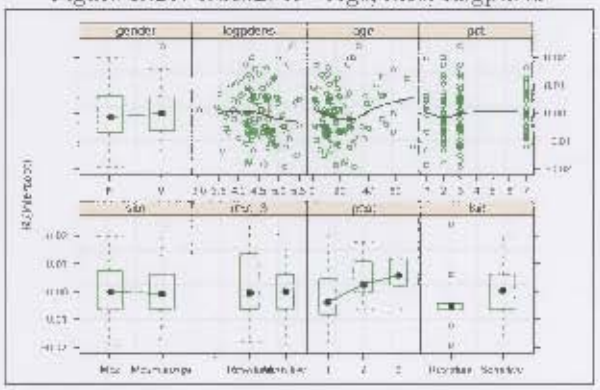


Figure B.22: Data2, C - Age (on R), Site, Logpdens, Peat

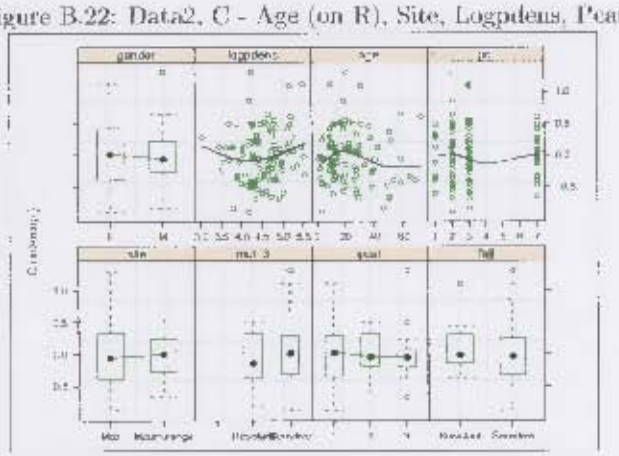


Figure B.23: Data2, R - Age, Site, Logpdens, Peat

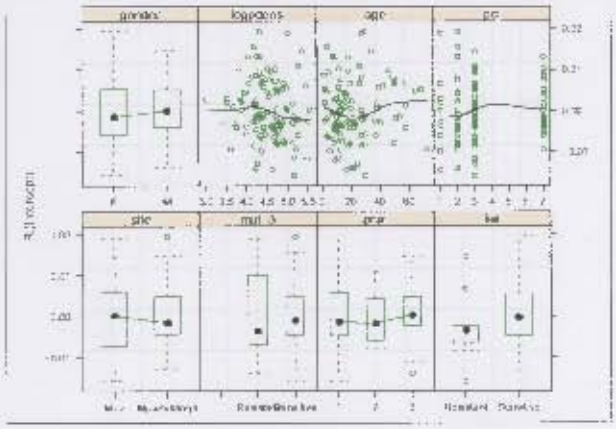


Figure B.24: Data2, Diagnostics - qqnorm

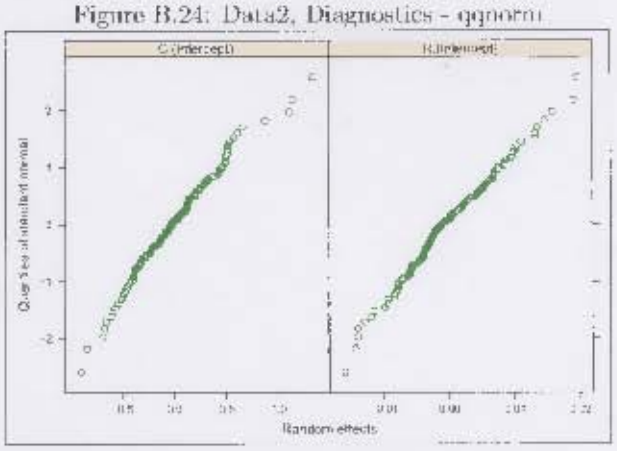


Figure B.25: Data2, Diagnostics - residuals

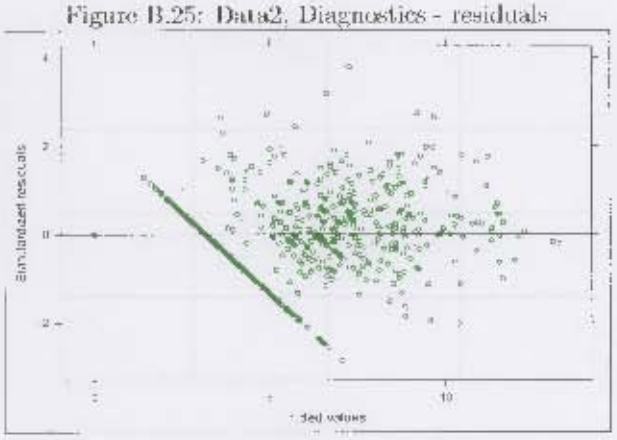
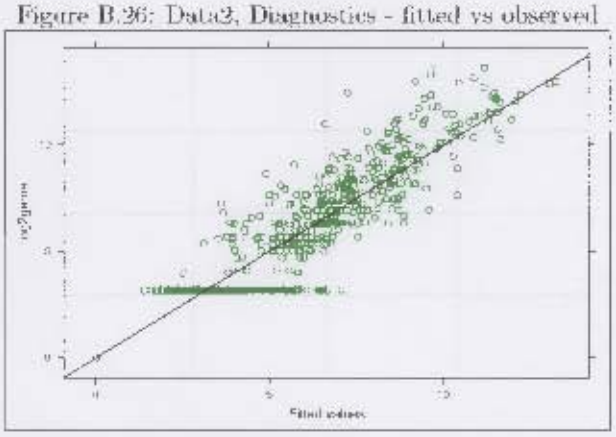


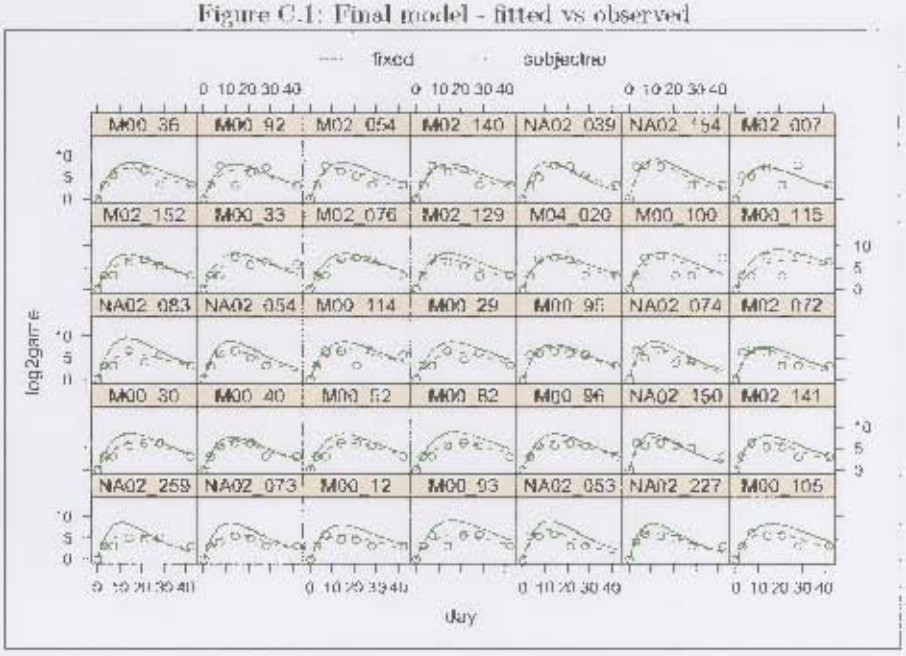
Figure B.26: Data3, Diagnostics - fitted vs observed

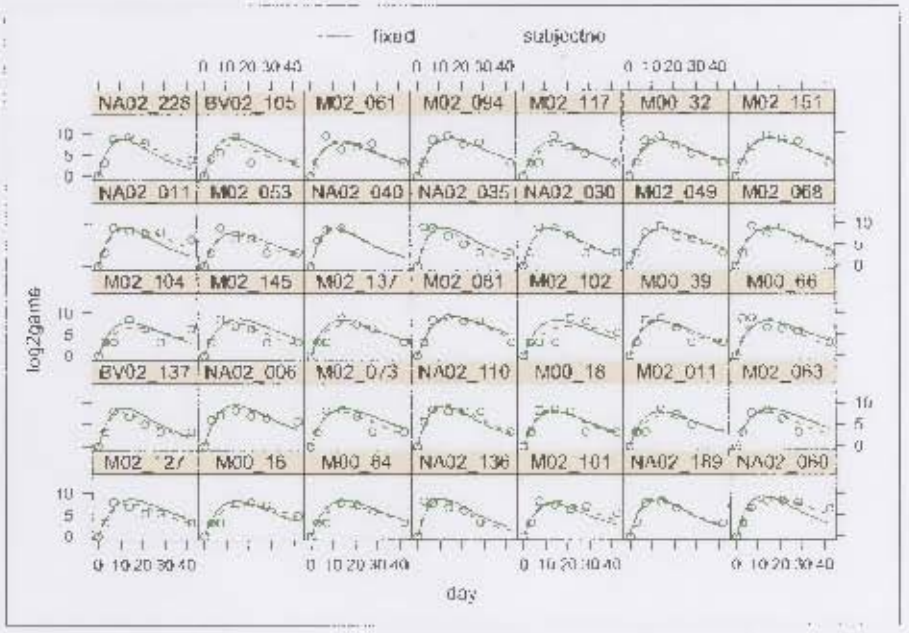


Appendix C

Model Fit Plots

Figure C.1: Final model - fitted vs observed




fixed
subjectno
NA02_228
BV02_105
M02_061
M02_094
M02_117
M00_32
M02_151
NA02_011
M02_053
NA02_040
NA02_035
NA02_030
M02_049
M02_068
M02_104
M02_145
M02_137
M02_081
M02_102
M00_39
M00_66
BV02_137
NA02_006
M02_073
NA02_110
M00_18
M02_011
M02_063
M02_27
M00_16
M00_84
NA02_136
M02_101
NA02_189
NA02_060
log2game
day

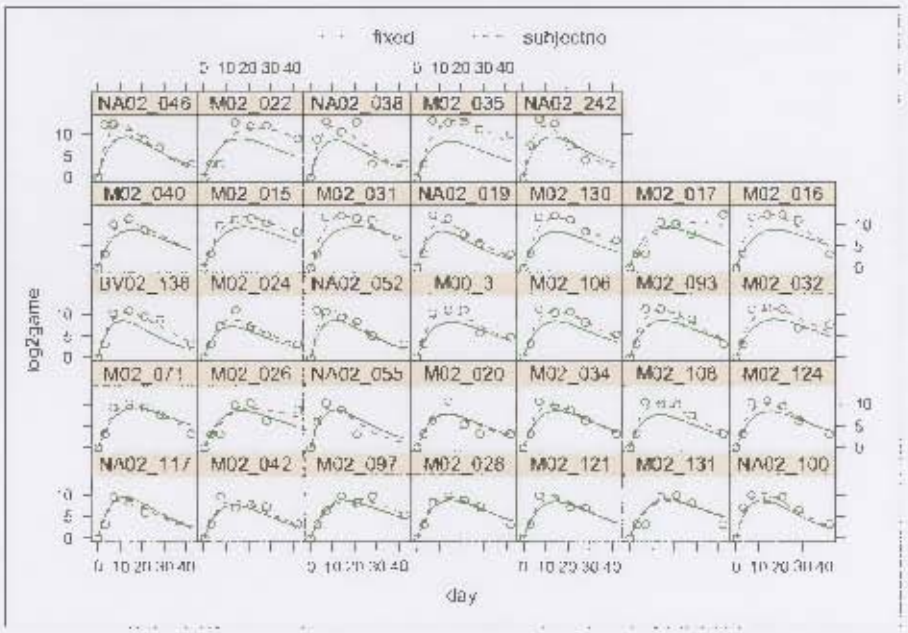
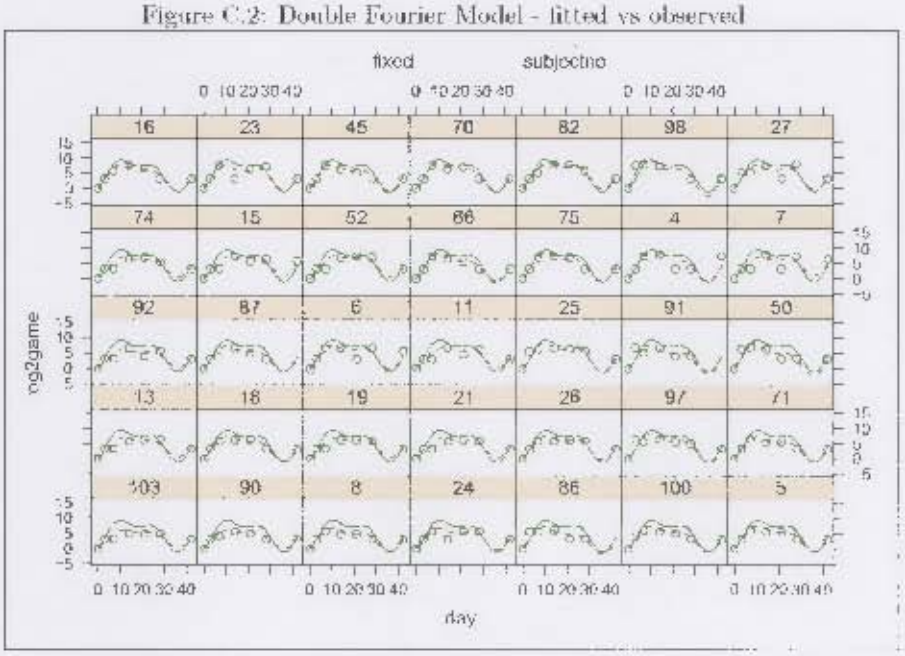
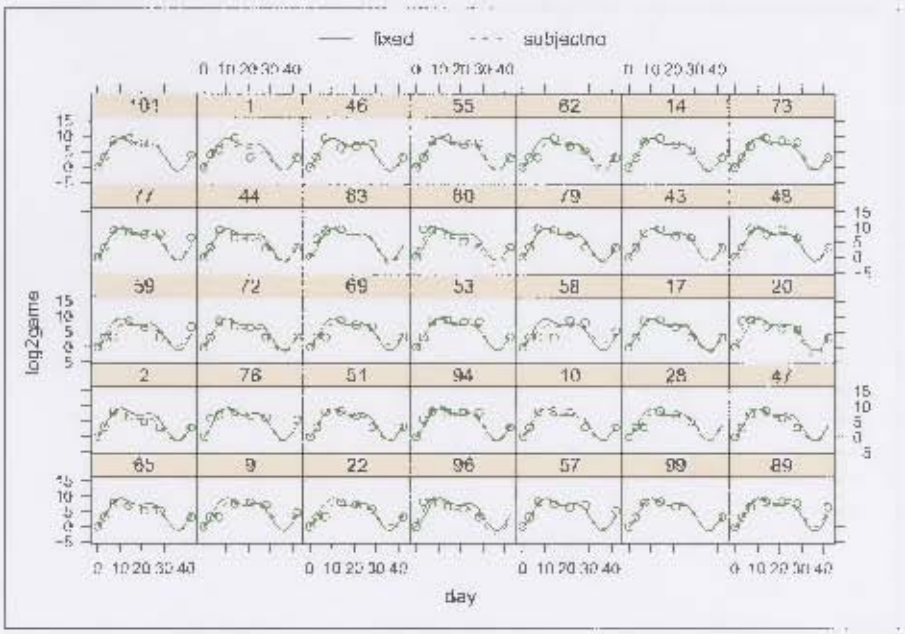
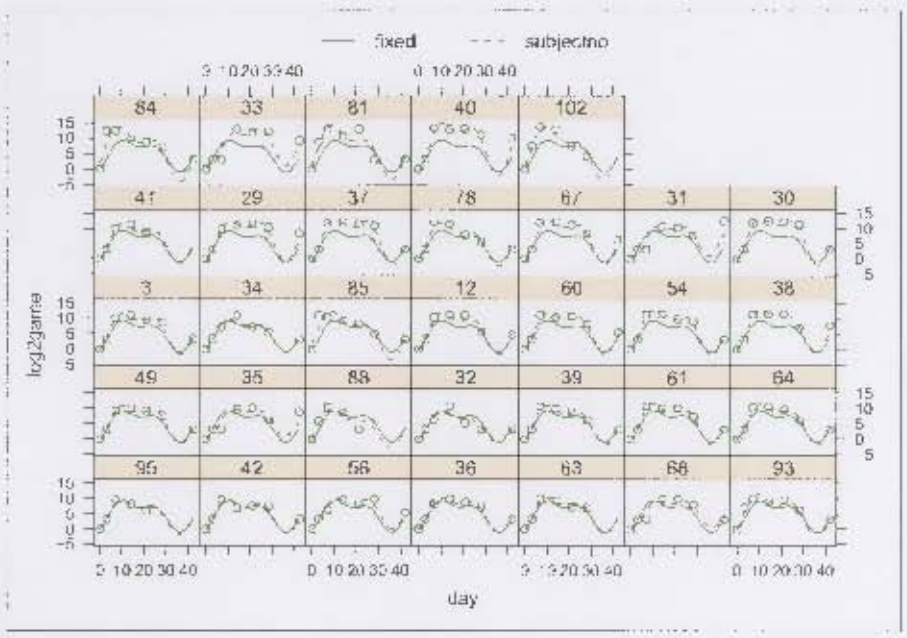

fixed
subjectno
NA02_046
M02_022
NA02_038
M02_035
NA02_242
M02_040
M02_015
M02_031
NA02_019
M02_130
M02_017
M02_016
BV02_138
M02_024
NA02_052
M00_3
M02_106
M02_093
M02_032
M02_071
M02_026
NA02_055
M02_020
M02_034
M02_108
M02_124
NA02_117
M02_042
M02_097
M02_028
M02_121
M02_131
NA02_100
log2game
day

Figure C.2: Double Fourier Model - fitted vs observed






fixed
subjectno
84
33
81
40
102
41
29
37
78
67
31
30
3
34
85
12
60
54
38
49
35
88
32
39
61
64
95
42
56
36
63
68
93
lcg12garnie
day

Appendix D

Computer Code

Stata 8.2 was used to prepare the data and perform the data exploration while R 2.2 was used for the actual modelling.

D.1 Stata 8.2

D.1.1 Data preparation

Preparing dataset for full sample with covariates

```
use "C:\Documents and Settings\gregd\My Documents\Msc\data\MBV2003SPASS.dta", clear
save "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta", replace
use "C:\Documents and Settings\gregd\My Documents\Msc\data\MBV2002SPASS.dta", clear
append using "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta"
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Moz_gam.dta", replace
use "C:\Documents and Settings\gregd\My Documents\Msc\data\MON2002SPASS.dta", clear
save "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta", replace
use "C:\Documents and Settings\gregd\My Documents\Msc\data\Moz_gam.dta", clear
append using "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta"
drop if day==1 | day==2
keep subjectno day pardens gamedens
reshape wide gamedens pardens, i(subjectno) j(day)
save "C:\Documents and Settings\gregd\My Documents\Msc\data\FullMoz.dta", replace

use "C:\Documents and Settings\gregd\My Documents\Msc\data\MPM19982004.dta", clear
drop if day==1
drop if day==2
keep subjectno day para game
rename para pardens
rename game gamedens
reshape wide gamedens pardens, i(subjectno) j(day)
append using "C:\Documents and Settings\gregd\My Documents\Msc\data\FullMoz.dta"
drop  pardens3 pardens7 pardens14 pardens21 pardens28 pardens42
sort subjectno
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Full.dta", replace

*So now have a file with all patients (game & pdens readings). Need covariate info.
use "C:\Documents and Settings\gregd\My Documents\Msc\data\MPM19982004.dta", clear
```

```
keep subjectno day para time out age gender pct fail dhfr dhps
reshape wide para, i(subjectno) j(day)
drop para1-para42
rename para0 pardens
gen site=2
label drop _all
sort subjectno
save "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta", replace
use "C:\Documents and Settings\gregd\My Documents\Msc\data\Full.dta", clear
merge subjectno using "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta"
sort subjectno
save "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", replace

*This just merges cov info for MPM
use "C:\Documents and Settings\gregd\My Documents\Msc\data\SP2002-3wide.dta", clear
drop if site==2
keep subjectno pardens time out age gender pct fail dhfrmax dhpsmax
label drop _all
gen site=1
sort subjectno
save "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta", replace
use "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", clear
drop _merge
merge subjectno using "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta" ,update
save "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", replace

use "C:\Documents and Settings\gregd\My Documents\Msc\data\MBV2003SPMUT.dta", clear
keep subjectno  dhfrmax dhpsmax
sort subjectno
save "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta", replace
use "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", clear
drop _merge
sort subjectno
merge subjectno using "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta" ,update
save "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", replace

use "C:\Documents and Settings\gregd\My Documents\Msc\data\Covdata.dta", clear
keep if subjectno=="MBV2002_034" | subjectno=="MBV2003_006" | subjectno=="MON2002_146"
keep subjectno pardens time out age gender pct fail dhfrmax dhpsmax
sort subjectno
save "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta", replace
use "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", clear
drop _merge
sort subjectno
merge subjectno using "C:\Documents and Settings\gregd\My Documents\Msc\data\temp.dta" ,update

gen mutcat1=0 if (dhfr=="w" & dhps=="W") | (dhfrmax==0 & dhpsmax==0)
replace mutcat1=5 if (dhfr=="t" & dhps=="D1") | (dhfrmax==3 & dhpsmax==2)
replace mutcat1=3 if (dhfr~="" & dhps~="" & mutcat==.) | (dhfrmax~=. & dhpsmax~=. & mutcat1==.)
gen mutcat2=mutcat1
replace mutcat2=0 if mutcat2==3
drop dhfr dhfrmax dhps dhpsmax
label define mutcat1 0 "Sensitive" 3 "Mixed" 5 "Resistant"
label value mutcat1 mutcat1
label value mutcat2 mutcat1
label define site 1 "Mozambique" 2 "Mpumulanga"
label value site site
```

```
label define fail 0 "Sensitive" 1 "Resistant" 2 "LTFU"
label value fail fail
drop out pardens _merge
*drops mpm2004 patients as they were on diff treatment, and 3 BV2003 poatients that were not in study.
drop in 582/600
drop in 583/657
drop if subjectno=="MBV2003_029" | subjectno=="MBV2003_031" | subjectno=="MBV2003_032"
save "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", replace

use "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", clear
reshape long gamedens, i(subjectno) j(day)
gen zero=(gamedens==0)
replace gamedens=. if gamedens==0
egen gamecount=count(gamedens),by(subjectno)
replace gamedens=0 if zero==1
drop zero
reshape wide gamedens, i(subjectno) j(day)
gen group=((gamedens0==0 | gamedens0==.) & (gamedens3==0 | gamedens3==.) & (gamedens7==0 | gamedens7==.) & (g
replace group=((gamedens0>0 & gamedens0~=.)|(gamecount<3)|(subjectno=="MPM1998_045" | subjectno=="MPM1998_090
replace group=(gamecount>2)*3 if group==0
save "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", replace
keep if group==3
gen logpdens=log10(pardens0)
*Manually add in pcat variable from excel file called Pcat.
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Grp3Cov.dta", replace
```

Preparing datasets for phase 1

```
*Construct Dataset1 with zeros made missing
use "C:\Documents and Settings\gregd\My Documents\Msc\data\Grp3Cov.dta", clear
reshape long gamedens, i(subjectno) j(day)
gen zero=(gamedens==0)
replace gamedens=8 if gamedens==1
replace gamedens=32 if gamedens==4
replace gamedens=. if gamedens==0
replace gamedens=0 if zero==1 & day==0
drop zero
drop if gamedens==.
sort subjectno day
preserve

save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_103.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_103.raw", comma replace
drop if pcat==3
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_88.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_88.raw", comma replace
drop if pcat==2
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_74.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_74.raw", comma replace

restore
keep if gamecount>3
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_56.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_56.raw", comma replace
drop if pcat==3
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_47.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_47.raw", comma replace
drop if pcat==2
```

```
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_34.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data1_34.raw", comma replace

*Construct Dataset2 with zeros changed to 8's
use "C:\Documents and Settings\gregd\My Documents\Msc\data\Grp3Cov.dta", clear
reshape long gamedens, i(subjectno) j(day)
gen zero=(gamedens==0)
replace gamedens=8 if gamedens==1
replace gamedens=32 if gamedens==4
replace gamedens=0 if zero==1 & day==0
replace gamedens=8 if zero==1 & day~=0
drop zero
drop if gamedens==.
sort subjectno day
preserve

save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_103.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_103.raw", comma replace
drop if pcat==3
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_88.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_88.raw", comma replace
drop if pcat==2
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_74.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_74.raw", comma replace

restore
keep if gamecount>3
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_56.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_56.raw", comma replace
drop if pcat==3
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_47.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_47.raw", comma replace
drop if pcat==2
save "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_34.dta", replace
outsheet using "C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Data5_34.raw", comma replace
```

D.1.2 Data exploration - Stata output

```
-------------------------------------------------------------------------------
      log:  C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Explor
> e.smcl
  log type:  smcl
 opened on:  18 Jan 2007, 09:08:03

. use "C:\Documents and Settings\gregd\My Documents\Msc\data\FullCov.dta", clear

. tab group

      group |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        305       52.68       52.68
          2 |        171       29.53       82.21
          3 |        103       17.79      100.00
------------+-----------------------------------
      Total |        579      100.00

. sort group
```

```
. by group: tab gender

-------------------------------------------------------------------------------
-> group = 1

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
          F |        166       54.43       54.43
          M |        139       45.57      100.00
------------+-----------------------------------
      Total |        305      100.00

-------------------------------------------------------------------------------
-> group = 2

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
          F |         83       48.54       48.54
          M |         88       51.46      100.00
------------+-----------------------------------
      Total |        171      100.00

-------------------------------------------------------------------------------
-> group = 3

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
          F |         53       51.46       51.46
          M |         50       48.54      100.00
------------+-----------------------------------
      Total |        103      100.00


. by group: tab site

-------------------------------------------------------------------------------
-> group = 1

       site |      Freq.     Percent        Cum.
------------+-----------------------------------
 Mozambique |         76       25.00       25.00
 Mpumulanga |        228       75.00      100.00
------------+-----------------------------------
      Total |        304      100.00

-------------------------------------------------------------------------------
-> group = 2

       site |      Freq.     Percent        Cum.
------------+-----------------------------------
 Mozambique |         65       38.46       38.46
 Mpumulanga |        104       61.54      100.00
------------+-----------------------------------
      Total |        169      100.00

-------------------------------------------------------------------------------
```

```
-> group = 3

       site |      Freq.     Percent        Cum.
------------+-----------------------------------
 Mozambique |         31       30.10       30.10
 Mpumulanga |         72       69.90      100.00
------------+-----------------------------------
      Total |        103      100.00

. by group: tab mutcat2

-------------------------------------------------------------------------------
-> group = 1

    mutcat2 |      Freq.     Percent        Cum.
------------+-----------------------------------
  Sensitive |        220       89.07       89.07
  Resistant |         27       10.93      100.00
------------+-----------------------------------
      Total |        247      100.00

-------------------------------------------------------------------------------
-> group = 2

    mutcat2 |      Freq.     Percent        Cum.
------------+-----------------------------------
  Sensitive |        128       84.21       84.21
  Resistant |         24       15.79      100.00
------------+-----------------------------------
      Total |        152      100.00

-------------------------------------------------------------------------------
-> group = 3

    mutcat2 |      Freq.     Percent        Cum.
------------+-----------------------------------
  Sensitive |         82       88.17       88.17
  Resistant |         11       11.83      100.00
------------+-----------------------------------
      Total |         93      100.00

. hist age
(bin=24, start=1, width=3.25)

. by group: swilk age

-------------------------------------------------------------------------------
-> group = 1

                  Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
         age |    305    0.89370     22.981      7.364  0.00000

-------------------------------------------------------------------------------
```

```
                            Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%             2             1
 5%             3             2
10%             5             2       Obs                 103
25%            10             2       Sum of Wgt.         103

50%            17                     Mean           21.75728
                        Largest       Std. Dev.      16.35481
75%            29            58
90%            50            64       Variance       267.4797
95%            58            68       Skewness       1.15426
99%            68            72       Kurtosis       3.666058


. by group: means  pardens0

-------------------------------------------------------------------------------
-> group = 1

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-----------------------------------------------------------------
    pardens0 | Arithmetic      304    37465.11         32701.8   42228.42
             |  Geometric      304    22629.94        20015.64    25585.7
             |   Harmonic      304    11374.07        9584.384   13985.58
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> group = 2

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-----------------------------------------------------------------
    pardens0 | Arithmetic      171    39507.23        31108.07    47906.4
             |  Geometric      171    23860.87        20348.25   27979.85
             |   Harmonic      171    12199.24        9602.892   16719.78
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> group = 3

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-----------------------------------------------------------------
    pardens0 | Arithmetic      103    56767.45        43627.19    69907.7
             |  Geometric      103    31641.89        25315.21   39549.72
             |   Harmonic      103    15825.35        11766.23   24160.09
-------------------------------------------------------------------------------

. by group: summ  pardens0

-------------------------------------------------------------------------------
-> group = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pardens0 |       304    37465.11    42204.54        1003     339333
```

```
-------------------------------------------------------------------------------
-> group = 2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pardens0 |       171    39507.23     55639.5        1008     598000

-------------------------------------------------------------------------------
-> group = 3

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pardens0 |       103    56767.45    67234.37        1040     332000


. by group: summ  pardens0 logpdens

-------------------------------------------------------------------------------
-> group = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pardens0 |       304    37465.11    42204.54        1003     339333
    logpdens |       304    4.354683    .4723794    3.001301   5.530626

-------------------------------------------------------------------------------
-> group = 2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pardens0 |       171    39507.23     55639.5        1008     598000
    logpdens |       171    4.377686    .4581385    3.003461   5.776701

-------------------------------------------------------------------------------
-> group = 3

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pardens0 |       103    56767.45    67234.37        1040     332000
    logpdens |       103    4.500262    .4957079    3.017033   5.521138


. tab pct

       pct |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |          1        0.18        0.18
         1 |         75       13.54       13.72
         2 |        276       49.82       63.54
         3 |        127       22.92       86.46
         7 |         75       13.54      100.00
-----------+-----------------------------------
     Total |        554      100.00

. replace pct=1 if pct==0
(1 real change made)
```

```
. by group: summ   pct,detail

-------------------------------------------------------------------------------
-> group = 1

                          pct
-------------------------------------------------------------
      Percentiles      Smallest
 1%           1             1
 5%           1             1
10%           1             1       Obs                 280
25%           2             1       Sum of Wgt.         280

50%           2                     Mean           2.421429
                      Largest       Std. Dev.      1.466803
75%           3             7
90%           3             7       Variance       2.15151
95%           7             7       Skewness       2.250671
99%           7             7       Kurtosis       7.667358

-------------------------------------------------------------------------------
-> group = 2

                          pct
-------------------------------------------------------------
      Percentiles      Smallest
 1%           1             1
 5%           1             1
10%           1             1       Obs                 171
25%           2             1       Sum of Wgt.         171

50%           2                     Mean           2.883041
                      Largest       Std. Dev.      1.875007
75%           3             7
90%           7             7       Variance       3.515652
95%           7             7       Skewness       1.512506
99%           7             7       Kurtosis       3.859268

-------------------------------------------------------------------------------
-> group = 3

                          pct
-------------------------------------------------------------
      Percentiles      Smallest
 1%           1             1
 5%           1             1
10%           2             1       Obs                 103
25%           2             1       Sum of Wgt.         103

50%           3                     Mean           3.524272
                      Largest       Std. Dev.      2.118289
75%           7             7
90%           7             7       Variance       4.48715
95%           7             7       Skewness       .8491875
99%           7             7       Kurtosis       2.146384

.preserve
```

```
. drop if fail==2
(35 observations deleted)

. by group: tab fail

-------------------------------------------------------------------------------------------
-> group = 1

       fail |      Freq.     Percent        Cum.
------------+-----------------------------------
  Sensitive |       236       85.51       85.51
  Resistant |        40       14.49      100.00
------------+-----------------------------------
      Total |       276      100.00

-------------------------------------------------------------------------------------------
-> group = 2

       fail |      Freq.     Percent        Cum.
------------+-----------------------------------
  Sensitive |       134       80.72       80.72
  Resistant |        32       19.28      100.00
------------+-----------------------------------
      Total |       166      100.00

-------------------------------------------------------------------------------------------
-> group = 3

       fail |      Freq.     Percent        Cum.
------------+-----------------------------------
  Sensitive |        93       91.18       91.18
  Resistant |         9        8.82      100.00
------------+-----------------------------------
      Total |       102      100.00

. tab group fail, exact col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |        fail
     group | Sensitive  Resistant |     Total
-----------+----------------------+----------
         1 |       236        40 |       276
           |     50.97      49.38 |     50.74
-----------+----------------------+----------
         2 |       134        32 |       166
           |     28.94      39.51 |     30.51
-----------+----------------------+----------
         3 |        93         9 |       102
           |     20.09      11.11 |     18.75
-----------+----------------------+----------
```

```
    Total |       463         81 |       544
          |    100.00     100.00 |    100.00

          Fisher's exact =                 0.064
```

```
. restore
```

```
. tab group gender,chi
```

```
           |          Sex
     group |         F          M |     Total
-----------+----------------------+----------
         1 |       166        139 |       305
         2 |        83         88 |       171
         3 |        53         50 |       103
-----------+----------------------+----------
     Total |       302        277 |       579

          Pearson chi2(2) =   1.5472   Pr = 0.461
```

```
. tab group site,chi row col
```

```
+--------------------+
| Key                |
|--------------------|
|     frequency      |
|  row percentage    |
| column percentage  |
+--------------------+
```

```
           |         site
     group | Mozambiqu  Mpumulang |     Total
-----------+----------------------+----------
         1 |        76        228 |       304
           |     25.00      75.00 |    100.00
           |     44.19      56.44 |     52.78
-----------+----------------------+----------
         2 |        65        104 |       169
           |     38.46      61.54 |    100.00
           |     37.79      25.74 |     29.34
-----------+----------------------+----------
         3 |        31         72 |       103
           |     30.10      69.90 |    100.00
           |     18.02      17.82 |     17.88
-----------+----------------------+----------
     Total |       172        404 |       576
           |     29.86      70.14 |    100.00
           |    100.00     100.00 |    100.00

          Pearson chi2(2) =   9.4011   Pr = 0.009
```

```
. tab mutcat2 group,chi col
```

```
+--------------------+
| Key                |
|--------------------|
|     frequency      |
```

```
| column percentage |
+-------------------+
```

mutcat2	group 1	2	3	Total
Sensitive	220	128	82	430
	89.07	84.21	88.17	87.40
Resistant	27	24	11	62
	10.93	15.79	11.83	12.60
Total	247	152	93	492
	100.00	100.00	100.00	100.00

Pearson chi2(2) = 2.0789 Pr = 0.354

```
. tab pct group,chi col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+
```

pct	group 1	2	3	Total
1	46	20	10	76
	16.43	11.70	9.71	13.72
2	158	88	30	276
	56.43	51.46	29.13	49.82
3	54	36	37	127
	19.29	21.05	35.92	22.92
7	22	27	26	75
	7.86	15.79	25.24	13.54
Total	280	171	103	554
	100.00	100.00	100.00	100.00

Pearson chi2(6) = 41.8246 Pr = 0.000

```
. graph box pct,medtype(line) by(group, rows(1)) ytitle (PCT)

. kwallis age,by(group)

Test: Equality of populations (Kruskal-Wallis test)

  +-----------------------+
  | group | Obs | Rank Sum |
  |-------+-----+----------|
  |     1 | 305 | 90298.50 |
```

```
  |     2 | 171 | 45147.00 |
  |     3 | 103 | 32464.50 |
  +------------------------+

chi-squared =     6.861 with 2 d.f.
probability =     0.0324

chi-squared with ties =     6.867 with 2 d.f.
probability =     0.0323

. graph box age,medtype(line) by(group, rows(1)) ytitle (Age)

. kwallis  pardens0,by(group)

Test: Equality of populations (Kruskal-Wallis test)

  +------------------------+
  | group | Obs | Rank Sum |
  |-------+-----+----------|
  |     1 | 304 | 85098.00 |
  |     2 | 171 | 48849.00 |
  |     3 | 103 | 33384.00 |
  +------------------------+

chi-squared =     5.515 with 2 d.f.
probability =     0.0635

chi-squared with ties =     5.515 with 2 d.f.
probability =     0.0635

. graph box  pardens0,medtype(line) by(group, rows(1)) ytitle (Parasite density Day 0
> )

. kwallis   logpdens,by(group)

Test: Equality of populations (Kruskal-Wallis test)

  +------------------------+
  | group | Obs | Rank Sum |
  |-------+-----+----------|
  |     1 | 304 | 85098.00 |
  |     2 | 171 | 48849.00 |
  |     3 | 103 | 33384.00 |
  +------------------------+

chi-squared =     5.515 with 2 d.f.
probability =     0.0635

chi-squared with ties =     5.515 with 2 d.f.
probability =     0.0635


. use "C:\Documents and Settings\gregd\My Documents\Msc\data\Grp3Cov.dta", clea
> r


. tab mutcat1
```

```
    mutcat1 |      Freq.     Percent        Cum.
------------+-----------------------------------
  Sensitive |         34       36.56       36.56
      Mixed |         48       51.61       88.17
  Resistant |         11       11.83      100.00
------------+-----------------------------------
      Total |         93      100.00
```

```
. tab fail

       fail |      Freq.     Percent        Cum.
------------+-----------------------------------
  Sensitive |         93       90.29       90.29
  Resistant |          9        8.74       99.03
       LTFU |          1        0.97      100.00
------------+-----------------------------------
      Total |        103      100.00

. replace gamedens=8 if gamedens>0 & gamedens~=.
(386 real changes made)

. preserve

. drop if day==0
(103 observations deleted)

. tab gamedens

   gamedens |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        211       35.34       35.34
          8 |        386       64.66      100.00
------------+-----------------------------------
      Total |        597      100.00

. restore

. reshape wide gamedens, i(subjectno) j(day)
(note: j = 0 3 7 14 21 28 42)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                      721   ->     103
Number of variables                  15   ->      20
j variable (7 values)               day   ->   (dropped)
xij variables:
                                gamedens   ->   gamedens0 gamedens3 ... gamedens
> 42
-----------------------------------------------------------------------------
```

```
. tab gamedens3

 3 gamedens |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         82       79.61       79.61
          8 |         21       20.39      100.00
------------+-----------------------------------
      Total |        103      100.00

. tab gamedens7

 7 gamedens |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         20       19.42       19.42
          8 |         83       80.58      100.00
------------+-----------------------------------
      Total |        103      100.00

. tab gamedens14

14 gamedens |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          3        2.91        2.91
          8 |        100       97.09      100.00
------------+-----------------------------------
      Total |        103      100.00

. tab gamedens21

21 gamedens |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          8        7.92        7.92
          8 |         93       92.08      100.00
------------+-----------------------------------
      Total |        101      100.00

. tab gamedens28

28 gamedens |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         29       30.21       30.21
          8 |         67       69.79      100.00
------------+-----------------------------------
      Total |         96      100.00

. tab gamedens42

42 gamedens |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         69       75.82       75.82
          8 |         22       24.18      100.00
------------+-----------------------------------
      Total |         91      100.00

. use "C:\Documents and Settings\gregd\My Documents\Msc\data\Grp3Cov.dta", clea
> r
```

```
. sort pcat

. tab pcat  mutcat1, exact

           |            mutcat1
      pcat | Sensitive      Mixed  Resistant |     Total
-----------+---------------------------------+----------
         1 |        24         36          7 |        67
         2 |         4          5          3 |        12
         3 |         6          7          1 |        14
-----------+---------------------------------+----------
     Total |        34         48         11 |        93

           Fisher's exact =                 0.680

. tab pcat  mutcat2, exact

           |       mutcat2
      pcat | Sensitive  Resistant |     Total
-----------+----------------------+----------
         1 |        60          7 |        67
         2 |         9          3 |        12
         3 |        13          1 |        14
-----------+----------------------+----------
     Total |        82         11 |        93

           Fisher's exact =                 0.303

. tab pcat  site, exact

           |        site
      pcat | Mozambiqu  Mpumulang |     Total
-----------+----------------------+----------
         1 |        24         50 |        74
         2 |         4         10 |        14
         3 |         3         12 |        15
-----------+----------------------+----------
     Total |        31         72 |       103

           Fisher's exact =                 0.700

. tab pcat  gender, exact

           |         Sex
      pcat |         F          M |     Total
-----------+----------------------+----------
         1 |        39         35 |        74
         2 |         7          7 |        14
         3 |         7          8 |        15
-----------+----------------------+----------
     Total |        53         50 |       103

           Fisher's exact =                 0.951

. tab pcat  pct, exact

           |                   pct
```

pcat	1	2	3	7	Total
1	6	23	27	18	74
2	3	5	2	4	14
3	1	2	8	4	15
Total	10	30	37	26	103

Fisher's exact = 0.272

```
. kwallis (pct), by(pcat)

Test: Equality of populations (Kruskal-Wallis test)
```

pcat	Obs	Rank Sum
1	74	3841.50
2	14	624.00
3	15	890.50

```
chi-squared =     1.778 with 2 d.f.
probability =     0.4111

chi-squared with ties =     1.950 with 2 d.f.
probability =     0.3773

. by pcat: summ age logpdens pct, detail

-------------------------------------------------------------------------------
-> pcat = 1

                            Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           1             1
 5%           3             2
10%           5             2       Obs                  74
25%          10             3       Sum of Wgt.          74

50%          18                     Mean            21.78378
                       Largest      Std. Dev.       15.81595
75%          29            58
90%          46            58       Variance        250.1444
95%          58            64       Skewness        1.139647
99%          72            72       Kurtosis        3.879378

                          logpdens
-------------------------------------------------------------
      Percentiles      Smallest
 1%     3.017033      3.017033
 5%     3.677607      3.525045
10%     3.861952       3.60206       Obs                  74
25%     4.221414      3.677607       Sum of Wgt.          74

50%     4.418952                     Mean            4.475745
```

		Largest	Std. Dev.	.4845985
75%	4.790144	5.263873		
90%	5.135705	5.438637	Variance	.2348357
95%	5.263873	5.456799	Skewness	-.1666981
99%	5.49785	5.49785	Kurtosis	3.173964

pct

	Percentiles	Smallest		
1%	1	1		
5%	1	1		
10%	2	1	Obs	74
25%	2	1	Sum of Wgt.	74
50%	3		Mean	3.5
		Largest	Std. Dev.	2.082214
75%	3	7		
90%	7	7	Variance	4.335616
95%	7	7	Skewness	.912073
99%	7	7	Kurtosis	2.255797

-> pcat = 2

Age

	Percentiles	Smallest		
1%	2	2		
5%	2	8		
10%	8	8	Obs	14
25%	9	9	Sum of Wgt.	14
50%	11		Mean	15.07143
		Largest	Std. Dev.	12.05414
75%	14	14		
90%	27	26	Variance	145.3022
95%	50	27	Skewness	1.914951
99%	50	50	Kurtosis	6.122788

logpdens

	Percentiles	Smallest		
1%	3.526339	3.526339		
5%	3.526339	3.775246		
10%	3.775246	3.884285	Obs	14
25%	4.0103	4.0103	Sum of Wgt.	14
50%	4.599645		Mean	4.474162
		Largest	Std. Dev.	.5143568
75%	4.737193	4.737193		
90%	5.110859	4.969495	Variance	.2645629
95%	5.243336	5.110859	Skewness	-.3590965
99%	5.243336	5.243336	Kurtosis	2.131489

pct

	Percentiles	Smallest

1%	1	1		
5%	1	1		
10%	1	1	Obs	14
25%	2	2	Sum of Wgt.	14
50%	2		Mean	3.357143
		Largest	Std. Dev.	2.468483
75%	7	7		
90%	7	7	Variance	6.093407
95%	7	7	Skewness	.7509031
99%	7	7	Kurtosis	1.81617

-> pcat = 3

Age

	Percentiles	Smallest		
1%	3	3		
5%	3	4		
10%	4	11	Obs	15
25%	12	12	Sum of Wgt.	15
50%	20		Mean	27.86667
		Largest	Std. Dev.	20.6116
75%	50	50		
90%	58	51	Variance	424.8381
95%	68	58	Skewness	.6062133
99%	68	68	Kurtosis	2.077537

logpdens

	Percentiles	Smallest		
1%	3.447158	3.447158		
5%	3.447158	3.926342		
10%	3.926342	4.234188	Obs	15
25%	4.298307	4.298307	Sum of Wgt.	15
50%	4.738979		Mean	4.645577
		Largest	Std. Dev.	.5418881
75%	4.950365	4.950365		
90%	5.341791	5.154485	Variance	.2936427
95%	5.521138	5.341791	Skewness	-.4783443
99%	5.521138	5.521138	Kurtosis	2.916572

pct

	Percentiles	Smallest		
1%	1	1		
5%	1	2		
10%	2	2	Obs	15
25%	3	3	Sum of Wgt.	15
50%	3		Mean	3.8
		Largest	Std. Dev.	2.077086
75%	7	7		
90%	7	7	Variance	4.314286

```
95%             7               7       Skewness        .7702843
99%             7               7       Kurtosis        2.07708


. graph box age, medtype(line) by(pcat, rows(1)) ytitle (Age)

. kwallis(age), by(pcat)

Test: Equality of populations (Kruskal-Wallis test)

  +----------------------+
  | pcat | Obs | Rank Sum |
  |------+-----+----------|
  |    1 |  74 |  3910.00 |
  |    2 |  14 |   526.50 |
  |    3 |  15 |   919.50 |
  +----------------------+

chi-squared =     4.760 with 2 d.f.
probability =     0.0925

chi-squared with ties =     4.767 with 2 d.f.
probability =     0.0922

. graph box logpdens, medtype(line) by(pcat, rows(1)) ytitle (Log Parasite Dens
> ity Day 0)

. kwallis(logpdens), by(pcat)

Test: Equality of populations (Kruskal-Wallis test)

  +----------------------+
  | pcat | Obs | Rank Sum |
  |------+-----+----------|
  |    1 |  74 |  3702.00 |
  |    2 |  14 |   725.00 |
  |    3 |  15 |   929.00 |
  +----------------------+

chi-squared =     1.981 with 2 d.f.
probability =     0.3713

chi-squared with ties =     1.981 with 2 d.f.
probability =     0.3713

. preserve

. drop if fail==2
(1 observation deleted)

. tab pcat  fail, exact

           |        fail
      pcat | Sensitive  Resistant |     Total
-----------+----------------------+----------
         1 |        66          7 |        73
         2 |        14          0 |        14
```

```
        3 |        13          2 |        15
----------+--------------------+----------
    Total |        93          9 |       102

          Fisher's exact =                 0.470

. restore

. mlogit pcat age, rrr

Iteration 0:   log likelihood = -81.308713
Iteration 4:   log likelihood =  -78.88549

Multinomial logistic regression                 Number of obs   =        103
                                                LR chi2(2)      =       4.85
                                                Prob > chi2     =     0.0886
Log likelihood =  -78.88549                     Pseudo R2       =     0.0298

------------------------------------------------------------------------------
        pcat |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
         age |   .9643435   .0242836    -1.44   0.149      .917904    1.013132
-------------+----------------------------------------------------------------
3            |
         age |   1.020123   .0160141     1.27   0.204     .9892142    1.051998
------------------------------------------------------------------------------
(Outcome pcat==1 is the comparison group)

. mlogit pcat  mutcat2, rrr

Iteration 0:   log likelihood = -73.051665
Iteration 4:   log likelihood = -72.030582

Multinomial logistic regression                 Number of obs   =         93
                                                LR chi2(2)      =       2.04
                                                Prob > chi2     =     0.3602
Log likelihood = -72.030582                     Pseudo R2       =     0.0140

------------------------------------------------------------------------------
        pcat |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
     mutcat2 |   1.233634   .1917447     1.35   0.177     .9096691    1.672974
-------------+----------------------------------------------------------------
3            |
     mutcat2 |   .9200723   .2046161    -0.37   0.708     .5950071    1.422728
------------------------------------------------------------------------------
(Outcome pcat==1 is the comparison group)

. mlogit pcat   site, rrr

Iteration 0:   log likelihood = -81.308713
Iteration 3:   log likelihood = -80.813377

Multinomial logistic regression                 Number of obs   =        103
                                                LR chi2(2)      =       0.99
```

```
                                                Prob > chi2     =     0.6094
Log likelihood = -80.813377                     Pseudo R2       =     0.0061

------------------------------------------------------------------------------
       pcat |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
2           |
       site |        1.2   .7699351     0.28   0.776      .341224    4.220101
------------+-----------------------------------------------------------------
3           |
       site |       1.92   1.327903     0.94   0.346     .4949901    7.447422
------------------------------------------------------------------------------
(Outcome pcat==1 is the comparison group)

. mlogit pcat  gender, rrr

Iteration 0:   log likelihood = -81.308713
Iteration 2:   log likelihood =  -81.21089

Multinomial logistic regression                 Number of obs   =       103
                                                LR chi2(2)      =      0.20
                                                Prob > chi2     =     0.9068
Log likelihood =  -81.21089                     Pseudo R2       =     0.0012

------------------------------------------------------------------------------
       pcat |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
2           |
      gender |   1.114286    .649665     0.19   0.853     .3554007    3.493613
------------+-----------------------------------------------------------------
3           |
      gender |   1.273469   .7227072     0.43   0.670      .418719    3.873061
------------------------------------------------------------------------------
(Outcome pcat==1 is the comparison group)

. mlogit pcat pct, rrr

Iteration 0:   log likelihood = -81.308713
Iteration 3:   log likelihood = -81.134596

Multinomial logistic regression                 Number of obs   =       103
                                                LR chi2(2)      =      0.35
                                                Prob > chi2     =     0.8402
Log likelihood = -81.134596                     Pseudo R2       =     0.0021

------------------------------------------------------------------------------
       pcat |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
2           |
         pct |   .9670151   .1381777    -0.23   0.814     .7308095    1.279565
------------+-----------------------------------------------------------------
3           |
         pct |   1.066748   .1381395     0.50   0.618     .8276275    1.374957
------------------------------------------------------------------------------
(Outcome pcat==1 is the comparison group)

. mlogit pcat logpdens, rrr
```

```
Iteration 0:   log likelihood = -81.308713
Iteration 3:   log likelihood = -80.526824

Multinomial logistic regression                 Number of obs   =        103
                                                LR chi2(2)      =       1.56
                                                Prob > chi2     =     0.4575
Log likelihood = -80.526824                     Pseudo R2       =     0.0096

------------------------------------------------------------------------------
        pcat |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
    logpdens |   .9934994   .5877904    -0.01   0.991     .3115773    3.167885
-------------+----------------------------------------------------------------
3            |
    logpdens |   2.072242   1.249485     1.21   0.227     .6356197    6.755904
------------------------------------------------------------------------------
(Outcome pcat==1 is the comparison group)

. mlogit pcat  mutcat2  site age gender pct logpdens, rrr

Iteration 0:   log likelihood = -73.051665
Iteration 4:   log likelihood = -68.190637

Multinomial logistic regression                 Number of obs   =         93
                                                LR chi2(12)     =       9.72
                                                Prob > chi2     =     0.6403
Log likelihood = -68.190637                     Pseudo R2       =     0.0665

------------------------------------------------------------------------------
        pcat |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
     mutcat2 |   1.394287   .2509213     1.85   0.065     .9798656    1.983982
        site |   2.609909    2.11101     1.19   0.236     .5347314     12.7384
         age |   .9572555   .0273971    -1.53   0.127     .9050365    1.012487
      gender |     .62164   .4161392    -0.71   0.478     .1673887    2.308616
         pct |   1.089164    .167995     0.55   0.580      .805012    1.473617
    logpdens |   1.262468   .8232902     0.36   0.721     .3516644    4.532235
-------------+----------------------------------------------------------------
3            |
     mutcat2 |   .9256551   .2173965    -0.33   0.742     .5841692    1.466762
        site |   1.862635   1.350878     0.86   0.391     .4495656    7.717247
         age |   1.007222   .0180141     0.40   0.687     .9725266    1.043155
      gender |   1.196648   .7330277     0.29   0.769     .3602038    3.975436
         pct |   1.015028   .1520803     0.10   0.921     .7567351    1.361484
    logpdens |   2.169446   1.366587     1.23   0.219     .6311843    7.456614
------------------------------------------------------------------------------
(Outcome pcat==1 is the comparison group)

. preserve

. drop if fail==2
(1 observation deleted)

. mlogit pcat fail, rrr
```

```
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

           |       mutcat2
      fail | Sensitive  Resistant |     Total
-----------+----------------------+----------
 Sensitive |        77          6 |        83
           |     95.06      54.55 |     90.22
-----------+----------------------+----------
 Resistant |         4          5 |         9
           |      4.94      45.45 |      9.78
-----------+----------------------+----------
     Total |        81         11 |        92
           |    100.00     100.00 |    100.00

          Fisher's exact =                 0.001
   1-sided Fisher's exact =                 0.001
```

```
. tab site fail, col exact
```

```
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

           |        fail
      site | Sensitive  Resistant |     Total
-----------+----------------------+----------
Mozambique |        26          4 |        30
           |     27.96      44.44 |     29.41
-----------+----------------------+----------
Mpumulanga |        67          5 |        72
           |     72.04      55.56 |     70.59
-----------+----------------------+----------
     Total |        93          9 |       102
           |    100.00     100.00 |    100.00

          Fisher's exact =                 0.443
   1-sided Fisher's exact =                 0.249
```

```
. by fail: summ age logpdens, detail

-------------------------------------------------------------------------------
-> fail = Sensitive

                            Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           1              1
 5%           3              2
10%           7              2       Obs                  93
25%          10              3       Sum of Wgt.          93
```

50%	18		Mean	22.35484
		Largest	Std. Dev.	16.76931
75%	29	58		
90%	50	64	Variance	281.2097
95%	58	68	Skewness	1.11825
99%	72	72	Kurtosis	3.466603

logpdens

	Percentiles	Smallest		
1%	3.017033	3.017033		
5%	3.60206	3.447158		
10%	3.790285	3.525045	Obs	93
25%	4.231623	3.526339	Sum of Wgt.	93
50%	4.426836		Mean	4.465285
		Largest	Std. Dev.	.4880819
75%	4.790144	5.263873		
90%	5.110859	5.438637	Variance	.2382239
95%	5.243336	5.456799	Skewness	-.2289029
99%	5.521138	5.521138	Kurtosis	3.004307

-> fail = Resistant

Age

	Percentiles	Smallest		
1%	2	2		
5%	2	4		
10%	2	9	Obs	9
25%	9	9	Sum of Wgt.	9
50%	12		Mean	16.11111
		Largest	Std. Dev.	11.55903
75%	27	19		
90%	33	27	Variance	133.6111
95%	33	30	Skewness	.2950278
99%	33	33	Kurtosis	1.56965

logpdens

	Percentiles	Smallest		
1%	4.062958	4.062958		
5%	4.062958	4.677004		
10%	4.062958	4.716337	Obs	9
25%	4.716337	4.719331	Sum of Wgt.	9
50%	4.816433		Mean	4.899418
		Largest	Std. Dev.	.4307737
75%	5.153815	5.109241		
90%	5.49785	5.153815	Variance	.185566
95%	5.49785	5.341791	Skewness	-.4680715
99%	5.49785	5.49785	Kurtosis	2.723393

```
. graph box age, medtype(line) by(fail, rows(1)) ytitle (Age)
```

```
. ranksum(age), by(fail)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        fail |      obs    rank sum    expected
-------------+---------------------------------
   Sensitive |       93        4872      4789.5
   Resistant |        9         381       463.5
-------------+---------------------------------
    combined |      102        5253        5253

unadjusted variance     7184.25
adjustment for ties      -10.40
                        ----------
adjusted variance       7173.85

Ho: age(fail==Sensitive) = age(fail==Resistant)
             z =   0.974
    Prob > |z| =   0.3300

. graph box  logpdens, medtype(line) by(fail, rows(1)) ytitle (Logged Parasite Density Day 0)

. ranksum( logpdens), by(fail)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        fail |      obs    rank sum    expected
-------------+---------------------------------
   Sensitive |       93        4583      4789.5
   Resistant |        9         670       463.5
-------------+---------------------------------
    combined |      102        5253        5253

unadjusted variance     7184.25
adjustment for ties       -0.04
                        ----------
adjusted variance       7184.21

Ho: logpdens(fail==Sensitive) = logpdens(fail==Resistant)
             z =  -2.436
    Prob > |z| =   0.0148

. graph box  pct, medtype(line) by(fail, rows(1)) ytitle (Pct)

. restore

. tab site mutcat2, exact row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |        mutcat2
```

```
      site | Sensitive  Resistant |     Total
-----------+----------------------+----------
Mozambique |        24          6 |        30
           |     80.00      20.00 |    100.00
-----------+----------------------+----------
Mpumulanga |        58          5 |        63
           |     92.06       7.94 |    100.00
-----------+----------------------+----------
     Total |        82         11 |        93
           |     88.17      11.83 |    100.00

          Fisher's exact =                 0.166
  1-sided Fisher's exact =                 0.093

. tab site pct, exact row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |                    pct
      site |         1          2          3          7 |     Total
-----------+--------------------------------------------+----------
Mozambique |         7         10         10          4 |        31
           |     22.58      32.26      32.26      12.90 |    100.00
-----------+--------------------------------------------+----------
Mpumulanga |         3         20         27         22 |        72
           |      4.17      27.78      37.50      30.56 |    100.00
-----------+--------------------------------------------+----------
     Total |        10         30         37         26 |       103
           |      9.71      29.13      35.92      25.24 |    100.00

          Fisher's exact =                 0.018

. ranksum(pct), by(site)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       site |      obs    rank sum    expected
------------+---------------------------------
 Mozambique |       31      1245.5        1612
 Mpumulanga |       72      4110.5        3744
------------+---------------------------------
   combined |      103        5356        5356

unadjusted variance     19344.00
adjustment for ties     -1701.83
                        ----------
adjusted variance       17642.17

Ho: pct(site==Mozambique) = pct(site==Mpumulanga)
            z =  -2.759
    Prob > |z| =   0.0058
```

```
. graph box  pct, medtype(line) by(site, rows(1)) ytitle (Pct)

. ranksum(age), by(site)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        site |      obs    rank sum    expected
-------------+---------------------------------
  Mozambique |       31        1417        1612
  Mpumulanga |       72        3939        3744
-------------+---------------------------------
    combined |      103        5356        5356

unadjusted variance    19344.00
adjustment for ties      -27.30
                       ----------
adjusted variance      19316.70

Ho: age(site==Mozambique) = age(site==Mpumulanga)
             z =  -1.403
    Prob > |z| =   0.1606

. graph box  age, medtype(line) by(site, rows(1)) ytitle (Age)

. ranksum(logpdens), by(site)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        site |      obs    rank sum    expected
-------------+---------------------------------
  Mozambique |       31        1501        1612
  Mpumulanga |       72        3855        3744
-------------+---------------------------------
    combined |      103        5356        5356

unadjusted variance    19344.00
adjustment for ties       -0.11
                       ----------
adjusted variance      19343.89

Ho: logpdens(site==Mozambique) = logpdens(site==Mpumulanga)
             z =  -0.798
    Prob > |z| =   0.4248

. graph box   logpdens, medtype(line) by(site, rows(1)) ytitle (Log Parasite De
> nsity Day 0)

. sort mutcat2

. graph box age, medtype(line) by( mutcat2, rows(1)) ytitle (Age)

. ranksum(age), by( mutcat2)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

     mutcat2 |      obs    rank sum    expected
-------------+---------------------------------
```

```
   Sensitive |       82      3748.5        3854
   Resistant |       11       622.5         517
-------------+---------------------------------
    combined |       93        4371        4371

unadjusted variance     7065.67
adjustment for ties      -10.38
                        ----------
adjusted variance       7055.28

Ho: age(mutcat2==Sensitive) = age(mutcat2==Resistant)
             z =  -1.256
    Prob > |z| =   0.2091

. graph box  logpdens, medtype(line) by( mutcat2, rows(1)) ytitle (Logged Paras
> ite Density Day 0)

. ranksum( logpdens), by( mutcat2)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

     mutcat2 |      obs    rank sum    expected
-------------+---------------------------------
   Sensitive |       82        3803        3854
   Resistant |       11         568         517
-------------+---------------------------------
    combined |       93        4371        4371

unadjusted variance     7065.67
adjustment for ties       -0.05
                        ----------
adjusted variance       7065.61

Ho: logpdens(mutcat2==Sensitive) = logpdens(mutcat2==Resistant)
             z =  -0.607
    Prob > |z| =   0.5440

. graph box age, medtype(line) by(pct, rows(1)) ytitle (Age)

. kwallis age, by(pct)

Test: Equality of populations (Kruskal-Wallis test)

  +----------------------+
  | pct | Obs | Rank Sum |
  |-----+-----+----------|
  |   1 |  10 |   440.50 |
  |   2 |  30 |  1388.00 |
  |   3 |  37 |  2123.50 |
  |   7 |  26 |  1404.00 |
  +----------------------+

chi-squared =     3.134 with 3 d.f.
probability =     0.3714

chi-squared with ties =     3.139 with 3 d.f.
probability =     0.3707
```

```
. spearman age pct

 Number of obs =     103
Spearman's rho =      0.1341

Test of Ho: age and pct are independent
    Prob > |t| =      0.1768

. spearman age  logpdens

 Number of obs =     103
Spearman's rho =      0.0805

Test of Ho: age and logpdens are independent
    Prob > |t| =      0.4190

. sort site

. by site: summ age logpdens, detail

-------------------------------------------------------------------------------
-> site = Mozambique

                          Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           1             1
 5%           2             2
10%           2             2       Obs                  31
25%           5             2       Sum of Wgt.          31

50%          17                     Mean            19.3871
                      Largest       Std. Dev.       17.12635
75%          29            46
90%          46            50       Variance        293.3118
95%          58            58       Skewness        1.078646
99%          64            64       Kurtosis        3.373348

                        logpdens
-------------------------------------------------------------
      Percentiles      Smallest
 1%     3.790285      3.790285
 5%     3.861952      3.861952
10%     3.975983      3.884285       Obs                  31
25%     4.186476      3.975983       Sum of Wgt.          31

50%     4.338337                     Mean            4.484966
                      Largest       Std. Dev.       .4450388
75%     4.782516      4.989957
90%     4.989957      5.153815       Variance        .1980595
95%     5.456799      5.456799       Skewness        .5685335
99%      5.49785       5.49785       Kurtosis        2.633107

-------------------------------------------------------------------------------
-> site = Mpumulanga
```

```
                            Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           3              3
 5%           7              3
10%           9              4       Obs                  72
25%          11              7       Sum of Wgt.          72

50%          18                      Mean            22.77778
                        Largest       Std. Dev.        16.0257
75%        28.5             58
90%          50             58       Variance        256.8232
95%          58             68       Skewness        1.239572
99%          72             72       Kurtosis        3.833784

                          logpdens
-------------------------------------------------------------
      Percentiles      Smallest
 1%     3.017033       3.017033
 5%     3.526339       3.447158
10%     3.760422       3.525045       Obs                  72
25%     4.247952       3.526339       Sum of Wgt.          72

50%     4.542368                      Mean            4.506848
                        Largest       Std. Dev.        .5188276
75%     4.841091       5.263873
90%     5.154485       5.341791       Variance         .269182
95%     5.263873       5.438637       Skewness        -.4361102
99%     5.521138       5.521138       Kurtosis        2.973994

. log close
      log:  C:\Documents and Settings\gregd\My Documents\Msc\data\Final\Explor
> e.smcl
  log type:  smcl
 closed on:  18 Jan 2007, 09:52:47
-------------------------------------------------------------------------------
```

D.2 R 2.2.0

D.2.1 Phase 1 - finding the structure

```
#########################################################################
########define functions########
###PK 1: 1 compartment with 2 parameterisations
PK1.1 <- function (X, beta0,beta1,beta2){
beta0 * ( exp(-beta1*X)-exp(-beta2*X)) }

PK1.2 <- function (X, beta0,beta1,beta2){
beta0 * ( exp(-exp(beta1)*X)-exp(-exp(beta2)*X))}

###PK 2: 2 compartment i.e. triexponential
PK2.1 <- function (X, A1, beta1, A2, beta2,A3,beta3){
(A1 * exp(-beta1*X))+ (A2*exp(-beta2*X)) + (A3*exp(-beta3*X)) }

###Genstat curves
DblEx <-function (X, A, B,R,C,S){
A + B*(R^X)+C*(S^X) }

CEx <- function (X,A,B,C,R) {
(A+(B+C*X)*R^X)}

CEx2 <- function (X,C,R) {
(C*X)*(R^X)}

Fourier <- function (X,A,B,E,W) {
A+B*sin(2*pi*(X-E)/W)}

DbFourier <- function (X,A,B,E,W,G,P){
A+B*sin(2*pi*(X-E)/W)+G*sin(4*pi*(X-P)/W)}

#########################################################################
########Read in data########
####Dataset 1 - only day0 zeros

Data<-read.csv("Data1_34.csv",header=T)
attach(Data)
log2game<- log2(gamedens+1)
Data1_34<-data.frame(cbind(subjectno,day,log2game))
Data1_34<-groupedData(log2game ~ day | subjectno, data=Data1_34)
detach(Data)

Data<-read.csv("Data1_47.csv",header=T)
attach(Data)
log2game<- log2(gamedens+1)
Data1_47<-data.frame(cbind(subjectno,day,log2game))
Data1_47<-groupedData(log2game ~ day | subjectno, data=Data1_47)
detach(Data)

Data<-read.csv("Data1_56.csv",header=T)
attach(Data)
log2game<- log2(gamedens+1)
Data1_56<-data.frame(cbind(subjectno,day,log2game))
Data1_56<-groupedData(log2game ~ day | subjectno, data=Data1_56)
detach(Data)
```

```
Data<-read.csv("Data1_74.csv",header=T)
attach(Data)
log2game<- log2(gamedens+1)
Data1_74<-data.frame(cbind(subjectno,day,log2game))
Data1_74<-groupedData(log2game ~ day | subjectno, data=Data1_74)
detach(Data)

Data<-read.csv("Data1_88.csv",header=T)
attach(Data)
log2game<- log2(gamedens+1)
Data1_88<-data.frame(cbind(subjectno,day,log2game))
Data1_88<-groupedData(log2game ~ day | subjectno, data=Data1_88)
detach(Data)

Data<-read.csv("Data1_103.csv",header=T)
attach(Data)
log2game<- log2(gamedens+1)
Data1_103<-data.frame(cbind(subjectno,day,log2game))
Data1_103<-groupedData(log2game ~ day | subjectno, data=Data1_103)
detach(Data)

#####################################################################
####Dataset 5 - day0 left as 0 and others changed to 8s

Data<-read.csv("Data5_34.csv",header=T)
attach(Data)
log2game<-log2(gamedens+1)
Data5_34<-data.frame(cbind(subjectno,day,log2game))
Data5_34<-groupedData(log2game ~ day | subjectno, data=Data5_34)
detach(Data)

Data<-read.csv("Data5_47.csv",header=T)
attach(Data)
log2game<-log2(gamedens+1)
Data5_47<-data.frame(cbind(subjectno,day,log2game))
Data5_47<-groupedData(log2game ~ day | subjectno, data=Data5_47)
detach(Data)

Data<-read.csv("Data5_56.csv",header=T)
attach(Data)
log2game<-log2(gamedens+1)
Data5_56<-data.frame(cbind(subjectno,day,log2game))
Data5_56<-groupedData(log2game ~ day | subjectno, data=Data5_56)
detach(Data)

Data<-read.csv("Data5_74.csv",header=T)
attach(Data)
log2game<-log2(gamedens+1)
Data5_74<-data.frame(cbind(subjectno,day,log2game))
Data5_74<-groupedData(log2game ~ day | subjectno, data=Data5_74)
detach(Data)

Data<-read.csv("Data5_88.csv",header=T)
attach(Data)
log2game<-log2(gamedens+1)
Data5_88<-data.frame(cbind(subjectno,day,log2game))
```

```
Data5_88<-groupedData(log2game ~ day | subjectno, data=Data5_88)
detach(Data)

Data<-read.csv("Data5_103.csv",header=T)
attach(Data)
log2game<-log2(gamedens+1)
Data5_103<-data.frame(cbind(subjectno,day,log2game))
Data5_103<-groupedData(log2game ~ day | subjectno, data=Data5_103)
detach(Data)

#########################################################################
#######Models######
iter<-nlmeControl(maxIter=500)

#####Dataset 1 - only day0 zeros
#####PK 1 comp PK1_model
##For 34
PK1_mod11.34nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_34,start=c(beta0=15,beta1=0.05,be
ta2=0.45),trace=T)
#Solves to 11.5, 0.016, 0.39
PK1_mod11.34lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_34,start=c(beta0=12,beta1=0.1
,beta2=0.3))
#Solves for 22 patients
PK1_mod11.34nlme <- nlme(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_34,fixed=(beta0+beta1+beta2~1),
random=(beta0+beta1+beta2~1),start=c(beta0=12,beta1=0.1,beta2=0.2))
intervals(PK1_mod11.34nlme)
#Solves without intervals!!
PK1_mod21.34nlme <-update(PK1_mod11.34nlme, random=(beta0+beta1~1))
intervals(PK1_mod21.34nlme)
#Solves with intervals!
anova(PK1_mod11.34nlme,PK1_mod21.34nlme)
#suggests that PK1_model with no ranef for beta2 better!
PK1_mod31.34nlme <-update(PK1_mod21.34nlme, random=pdBlocked(list(beta2~1,beta0+beta1~1)),control=iter)
#solves with intervals
anova(PK1_mod31.34nlme,PK1_mod21.34nlme)
#PK1_mod 2 better
summary(PK1_mod21.34nlme)

##For 47
PK1_mod11.47nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_47,start=c(beta0=12,beta1=0.1,bet
a2=0.3))
#Solves to: 11.1, 0.014, 0.407
PK1_mod11.47lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_47,start=c(beta0=12,beta1=0.1
,beta2=0.2))
#Solves for 25
PK1_mod11.47nlme <- nlme(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_47,fixed=(beta0+beta1+beta2~1),
random=(beta0+beta1+beta2~1),start=c(beta0=12,beta1=0.1,beta2=0.2))
intervals(PK1_mod11.47nlme)
#Solves but no intervals, ranef for b2 small, also no corr btwn b2 and others
PK1_mod21.47nlme <-update(PK1_mod11.47nlme, random=(beta0+beta1~1))
intervals(PK1_mod21.47nlme)
#Solves with intervals
anova(PK1_mod11.47nlme,PK1_mod21.47nlme)
#PK1_mod 2 better
PK1_mod31.47nlme <-update(PK1_mod21.47nlme, random=pdBlocked(list(beta2~1,beta0+beta1~1)),control=iter)
#does not solve
summary(PK1_mod21.47nlme)
```

```
##For 56
PK1_mod11.56nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_56,start=c(beta0=12,beta1=0.1,bet
a2=0.3))
#Solves to: 10.5, 0.012, 0.42
PK1_mod11.56lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_56,start=c(beta0=12,beta1=0.1
,beta2=0.3))
#Solves for 27
PK1_mod11.56nlme <- nlme(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_56,fixed=(beta0+beta1+beta2~1),
random=(beta0+beta1+beta2~1),start=c(beta0=12,beta1=0.1,beta2=0.2))
intervals(PK1_mod11.56nlme)
#Solves without intervals
PK1_mod21.56nlme <-update(PK1_mod11.56nlme, random=(beta0+beta1~1))
intervals(PK1_mod21.56nlme)
#Solves with intervals
anova(PK1_mod11.56nlme,PK1_mod21.56nlme)
#PK1_mod 2 better
PK1_mod31.56nlme <-update(PK1_mod21.56nlme, random=pdBlocked(list(beta2~1,beta0+beta1~1)),control=iter)
#Cannot solve, singular precision
PK1_mod41.56nlme <-update(PK1_mod21.56nlme, random=pdDiag(beta0+beta1~1),control=iter)
anova(PK1_mod41.56nlme,PK1_mod21.56nlme)
#PK1_mod 2 better
summary(PK1_mod21.56nlme)

##For 74
PK1_mod11.74nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_74,start=c(beta0=12,beta1=0.1,bet
a2=0.2),trace=T)
#Solves to: 9.84, 0.013, 0.453
PK1_mod11.74lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_74,start=c(beta0=12,beta1=0.1
,beta2=0.2))
#Solves for 22
PK1_mod11.74nlme <- nlme(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_74,fixed=(beta0+beta1+beta2~1),
random=(beta0+beta1+beta2~1),start=c(beta0=12,beta1=0.1,beta2=0.2),verbose=T,control=iter)
intervals(PK1_mod11.74nlme)
#Solves but no intervals
PK1_mod21.74nlme <-nlme(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_74,fixed=(beta0+beta1+beta2~1),r
andom=(beta0+beta1~1),start=c(beta0=12,beta1=0.1,beta2=0.2),verbose=T)
intervals(PK1_mod21.74nlme)
#Solves with intervals
anova(PK1_mod21.74nlme,PK1_mod11.74nlme)
#PK1_mod 2 better!
PK1_mod31.74nlme <-update(PK1_mod21.74nlme, random=pdBlocked(list(beta2~1,beta0+beta1~1)),control=iter)
intervals(PK1_mod31.74nlme)
#Solves without intervals
anova(PK1_mod31.74nlme,PK1_mod21.74nlme)
#PK1_mod 2 better
summary(PK1_mod21.74nlme)

##For 88
PK1_mod11.88nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_88,start=c(beta0=12,beta1=0.1,bet
a2=0.3))
#Solves to: 9.7, 0.011, 0.464
PK1_mod11.88lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_88,start=c(beta0=12,beta1=0.1
,beta2=0.3))
#Solves for 25
PK1_mod11.88nlme <- nlme(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_88,fixed=(beta0+beta1+beta2~1),
random=(beta0+beta1+beta2~1),start=c(beta0=12,beta1=0.1,beta2=0.2),verbose=T)
```

```
intervals(PK1_mod11.88nlme)
#Solves but no intervals
PK1_mod21.88nlme <- update(PK1_mod11.88nlme,fixed=(beta0+beta1+beta2~1),random=(beta0+beta1~1),start=c(beta
0=12,beta1=0.1,beta2=0.2),verbose=T)
intervals(PK1_mod21.88nlme)
#Solves with intervals
anova(PK1_mod21.88nlme,PK1_mod11.88nlme)
#PK1_mod 2 better
PK1_mod31.88nlme <-update(PK1_mod21.88nlme, random=pdBlocked(list(beta2~1,beta0+beta1~1)),control=iter)
intervals(PK1_mod31.88nlme)
#Cannot solve, maxiter
summary(PK1_mod21.88nlme)

##For 103
PK1_mod11.103nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_103,start=c(beta0=12,beta1=0.1,b
eta2=0.3))
#Solves to 9.3, 0.01, 0.49
PK1_mod11.103lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_103,start=c(beta0=12,beta1=0
.1,beta2=0.3))
#Solves for 27 patients
PK1_mod11.103nlme <- nlme(PK1_mod11.103lis,random=(beta0+beta1+beta2~1),verbose=T)
#Cannot fit to the lis object! max iter
PK1_mod11.103nlme <- nlme(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data1_103,fixed=(beta0+beta1+beta2~1),
intervals(PK1_mod11.103nlme)
#Solves with intervals! Problem with beta2
PK1_mod21.103nlme <-update(PK1_mod11.103nlme, random=(beta0+beta1~1))
intervals(PK1_mod21.103nlme)
#Solves with intervals!!
anova(PK1_mod11.103nlme,PK1_mod21.103nlme)
#suggests that PK1_model with no ranef for beta2 better!
PK1_mod31.103nlme <-update(PK1_mod21.103nlme, random=pdBlocked(list(beta2~1,beta0+beta1~1)),control=iter)
intervals(PK1_mod31.103nlme)
#Solves with intervals
anova(PK1_mod31.103nlme,PK1_mod21.103nlme)
#best model is block diag

summary(PK1_mod31.103nlme)

##########Biexponential##################
##For 34
BI_mod11.34nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_34,trace=T)
#Solves to: -11.5, -0.93, 11.5, -4.09
BI_mod11.34lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_34)
#Solves for 27
plot(intervals(BI_mod11.34lis))
BI_mod11.34nlme <- nlme(BI_mod11.34lis,random=(A1+lrc1+A2+lrc2~1),control=iter)
#Cannot solve, max iter
BI_mod21.34nlme <- nlme(BI_mod11.34lis,random=pdDiag(A1+lrc1+A2+lrc2~1),control=iter)
intervals(BI_mod21.34nlme)
#Solves but no intervals, ranef for A1 and A2 v small
BI_mod31.34nlme <- update(BI_mod21.34nlme,random=pdDiag(lrc1+lrc2~1),control=iter)
intervals(BI_mod31.34nlme)
#Solves with intervals
anova(BI_mod21.34nlme,BI_mod31.34nlme)
#BI_mod 3 better
BI_mod41.34nlme <- update(BI_mod21.34nlme,random=(lrc1+lrc2~1),control=iter)
#Does not solve, singularity
```

```
summary(BI_mod31.34nlme)

##For 47
BI_mod11.47nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_47,trace=T)
#Solves to: -11.08, -0.9, 11.09, -4.27
BI_mod11.47lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_47)
#Solved for 34
BI_mod11.47nlme <- nlme(BI_mod11.47lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, max iter
BI_mod21.47nlme <- nlme(BI_mod11.47lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T)
intervals(BI_mod21.47nlme)
#Solves but no intervals, ranef for A1, A2 v small
BI_mod31.47nlme <- nlme(BI_mod11.47lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
intervals(BI_mod31.47nlme)
#Solves with intervals!
anova(BI_mod21.47nlme,BI_mod31.47nlme)
#BI_mod 3 better
BI_mod41.47nlme <- nlme(BI_mod11.47lis,random=(lrc1+lrc2~1),control=iter)
#Does not solve, step halving
summary(BI_mod31.47nlme)

##For 56
BI_mod11.56nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_56,trace=T)
#Solves to: -10.49, -0.86, 10.5, -4.45
BI_mod11.56lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_56)
#Solved for 36
BI_mod11.56nlme <- nlme(BI_mod11.56lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, max iter
BI_mod21.56nlme <- nlme(BI_mod11.56lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving
BI_mod31.56nlme <- nlme(BI_mod11.56lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving

##For 74
BI_mod11.74nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_74,trace=T)
#Solves to: -9.8, -0.79, 9.84, -4.32
BI_mod11.74lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_74)
#Solves for 27
BI_mod11.74nlme <- nlme(BI_mod11.74lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, max iter
BI_mod21.74nlme <- nlme(BI_mod11.74lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving
BI_mod31.74nlme <- nlme(BI_mod11.74lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving

##For 88
BI_mod11.88nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_88,trace=T)
#Solves to: -9.7, -0.77, 9.7, -4.48
BI_mod11.88lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_88)
#Solved for 34
BI_mod11.88nlme <- nlme(BI_mod11.88lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, singular precision
BI_mod21.88nlme <- nlme(BI_mod11.88lis,random=pdDiag(A1+lrc1+A2+lrc2~1),control=iter)
#Cannot solve, step halving
BI_mod31.88nlme <- nlme(BI_mod11.88lis,random=pdDiag(lrc1+lrc2~1),control=iter)
#Cannot solve, step halving
BI_mod41.88nlme <- nlme(BI_mod11.88lis,random=(lrc1+lrc2~1),control=iter)
```

```
#Cannot solve, step halving

##For 103
BI_mod11.103nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_103,trace=T)
#Solves to: -9.32, -0.72, 9.32, -4.66
BI_mod11.103lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data1_103)
#Solved for 36
BI_mod11.103nlme <- nlme(BI_mod11.103lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, singularity in backsolve...
BI_mod21.103nlme <- nlme(BI_mod11.103lis,random=pdDiag(A1+lrc1+A2+lrc2~1),control=iter)
#Cannot solve, step halving
BI_mod31.103nlme <- nlme(BI_mod11.103lis,random=pdDiag(lrc1+lrc2~1),control=iter)
#Cannot solve, step halving
BI_mod41.103nlme <- nlme(BI_mod11.103lis,random=(lrc1+lrc2~1),control=iter)
#Cannot solve, step halving

#########Triexponential#########
##For 34
PK2_mod11.34nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data1_34,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! singular gradient

##For 47
PK2_mod11.47nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data1_47,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! step factor / singular gradient

##For 56
PK2_mod11.56nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data1_56,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! step factor / singular gradient

##For 74
PK2_mod11.74nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data1_74,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! step factor / singular gradient

##For 88
PK2_mod11.88nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data1_88,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! step factor / singular gradient

##For 103
PK2_mod11.103nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data1_103,start=c(A1=13,beta1
=0.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! step factor / singular gradient

##########C exponential#########
##For 34
CEX_mod11.34nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data1_34,start=c(A=6.8,B=-5.5,C=2.5,R=0.9),trace=
T)
#Solves to: 6.11, -5.99, 2.24, 0.86
CEX_mod11.34lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data1_34,start=c(A=6.8,B=-5.5,C=2.5,R=0.9))
#Solved for 32
plot(intervals(CEX_mod11.34lis))
CEX_mod11.34nlme<-nlme(CEX_mod11.34lis,random=A+B+C+R~1,verbose=T,control=iter)
#Cannot solve, max iter
```

```
CEX_mod21.34nlme<-nlme(CEX_mod11.34lis,random=pdDiag(A+B+C+R^1),verbose=T,control=iter)
intervals(CEX_mod21.34nlme)
#Solves without intervals, A and B ranef v small
CEX_mod31.34nlme<-update(CEX_mod21.34nlme,random=pdDiag(C+R^1),control=iter)
intervals(CEX_mod31.34nlme)
#Solves with intervals
anova(CEX_mod31.34nlme,CEX_mod21.34nlme)
#CEX_mod 3 better
CEX_mod41.34nlme<-update(CEX_mod21.34nlme,random=(C+R^1),control=iter)
#Cannot solve, max iter
anova(CEX_mod31.34nlme,CEX_mod11.34nls)
#NLME fit better than NLS
summary(CEX_mod31.34nlme)

##For 47
CEX_mod11.47nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data1_47,start=c(A=7,B=-6,C=2.5,R=0.9),trace=T)
#Solves to: 6.63, -6.55, 2.18, 0.85
CEX_mod11.47lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data1_47,start=c(A=7,B=-6,C=2.5,R=0.9))
#Solved for 41
plot(intervals(CEX_mod11.47lis))
CEX_mod11.47nlme<-nlme(CEX_mod11.47lis,random=A+B+C+R^1,verbose=T,control=iter)
#Cannot solve, max iter
CEX_mod21.47nlme<-nlme(CEX_mod11.47lis,random=pdDiag(A+B+C+R^1),verbose=T,control=iter)
#Cannot solve, max iter
CEX_mod31.47nlme<-nlme(CEX_mod11.47lis,random=pdDiag(C+R^1),control=iter)
#Cannot solve, max iter
CEX_mod41.47nlme<-nlme(CEX_mod11.47lis,random=(C+R^1),control=iter)
#Cannot solve, max iter
CEX_mod51.47nlme<-nlme(CEX_mod11.47lis,random=(R^1),control=iter)
#Cannot solve, max iter
CEX_mod61.47nlme<-nlme(CEX_mod11.47lis,random(C^1),control=iter)
intervals(CEX_mod61.47nlme)
#Solves with intervals
summary((CEX_mod61.47nlme)

##For 56
CEX_mod11.56nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data1_56,start=c(A=7.5,B=-6.2,C=2.2,R=0.93),trace
=T)
#Solves to: 6.83, -6.74, 2.03, 0.85
CEX_mod11.56lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data1_56,start=c(A=7.3,B=-6,C=2.2,R=0.93))
#Solved for 45
plot(intervals(CEX_mod11.56lis))
CEX_mod11.56nlme<-nlme(CEX_mod11.56lis,random=A+B+C+R^1,verbose=T,control=iter)
#Cannot solve, max iter
CEX_mod21.56nlme<-nlme(CEX_mod11.56lis,random=pdDiag(A+B+C+R^1),verbose=T,control=iter)
#Cannot solve, max iter
CEX_mod31.56nlme<-nlme(CEX_mod11.56lis,random=(C+R^1),control=iter)
#Cannot solve, max iter
CEX_mod41.56nlme<-nlme(CEX_mod11.56lis,random=pdDiag(C+R^1),verbose=T,control=iter)
#Cannot solve, max iter
CEX_mod51.56nlme<-nlme(CEX_mod11.56lis,random=(R^1),verbose=T,control=iter)
intervals(CEX_mod51.56nlme)
#Solves with intervals
CEX_mod61.56nlme<-nlme(CEX_mod11.56lis,random=(C^1),control=iter)
#Cannot solve, max iter
anova(CEX_mod51.56nlme,CEX_mod11.56nls)
#NLS fit better!
```

```
##For 74
CEX_mod11.74nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data1_74,start=c(A=7,B=-6,C=2.5,R=0.9),trace=T)
#Solves to: 6.35, -6.29, 2.02, 0.84
CEX_mod11.74lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data1_74,start=c(A=7,B=-6,C=2.2,R=0.9))
#Solved for 32
plot(intervals(CEX_mod11.74lis))
CEX_mod11.74nlme<-nlme(CEX_mod11.74lis,random=A+B+C+R~1,verbose=T,control=iter)
#Cannot solve, max iter
CEX_mod21.74nlme<-nlme(CEX_mod11.74lis,random=pdDiag(A+B+C+R~1),verbose=T,control=iter)
intervals(CEX_mod21.74nlme)
#Solves without intervals
CEX_mod31.74nlme<-nlme(CEX_mod11.74lis,random=(C+R~1),control=iter)
#Cannot solve, max iter
CEX_mod41.74nlme<-nlme(CEX_mod11.74lis,random=pdDiag(C+R~1),control=iter)
#Cannot solve, max iter
CEX_mod51.74nlme<-nlme(CEX_mod11.74lis,random=(R~1),control=iter)
intervals(CEX_mod51.74nlme)
#Solves with intervals
CEX_mod61.74nlme<-nlme(CEX_mod11.74lis,random=(C~1),control=iter)
intervals(CEX_mod61.74nlme)
#Solves with intervals
anova(CEX_mod51.74nlme,CEX_mod61.74nlme)
#much better on C only
anova(anova(CEX_mod41.74nlme,CEX_mod11.74nls),CEX_mod61.74nlme)
#Mod 4 better i.e. with diag
anova(CEX_mod41.74nlme,CEX_mod11.74nls)
#better than NLS, Mod 4 best!
Summary(CEX_mod41.74nlme)

##For 88
CEX_mod11.88nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data1_88,start=c(A=7,B=-6,C=2.5,R=0.9),trace=T)
#Solves to: 6.7, -6.67, 2, 0.83
CEX_mod11.88lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data1_88,start=c(A=7,B=-6,C=2.5,R=0.9))
#Solved for 41
plot(intervals(CEX_mod11.88lis))
CEX_mod11.88nlme<-nlme(CEX_mod11.88lis,random=A+B+C+R~1,verbose=T,control=iter)
#Cannot solve, max iter
CEX_mod21.88nlme<-nlme(CEX_mod11.88lis,random=pdDiag(A+B+C+R~1),verbose=T,control=iter)
intervals(CEX_mod21.88nlme)
#Solves without intervals, all ranef v small
CEX_mod31.88nlme<-update(CEX_mod21.88nlme,random=(C+R~1))
#Cannot solve, max iter
CEX_mod41.88nlme<-nlme(CEX_mod11.88lis,random=pdDiag(C+R~1),verbose=T,control=iter)
intervals(CEX_mod41.88nlme)
#Solves without intervals
anova(CEX_mod21.88nlme,CEX_mod41.88nlme)
#CEX_mod 4 better
CEX_mod51.88nlme<-update(CEX_mod21.88nlme,random=(R~1))
intervals(CEX_mod51.88nlme)
#Solves without intervals
CEX_mod61.88nlme<-nlme(CEX_mod11.88lis,random=pdDiag(C~1),verbose=T,control=iter)
intervals(CEX_mod61.88nlme)
#Solves with intervals
anova(CEX_mod51.88nlme,CEX_mod61.88nlme)
anova(CEX_mod41.88nlme,CEX_mod61.88nlme)
#mod with 1 ranef better
```

```
anova(CEX_mod61.88nlme,CEX_mod11.88nls)
#Nls fit better!

##For 103
CEX_mod11.103nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data1_103,start=c(A=7.5,B=-6.2,C=2.2,R=0.93),tra
ce=T)
#Solves to: 6.79, -6.75, 1.92, 0.83
CEX_mod11.103lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data1_103,start=c(A=7.5,B=-6.2,C=2.2,R=0.9))
#Solved for 47
plot(intervals(CEX_mod11.103lis))
CEX_mod11.103nlme<-nlme(CEX_mod11.103lis,random=A+B+C+R~1,verbose=T,control=iter)
#Cannot solve, max iter
CEX_mod21.103nlme<-nlme(CEX_mod11.103lis,random=pdDiag(A+B+C+R~1),verbose=T,control=iter)
intervals(CEX_mod21.103nlme)
#Solves without intervals, all ranef v small
CEX_mod31.103nlme<-update(CEX_mod21.103nlme,random=(C+R~1),control=iter)
#Cannot solve, max iter
CEX_mod41.103nlme<-update(CEX_mod21.103nlme,random=pdDiag(C+R~1),control=iter)
intervals(CEX_mod41.103nlme)
#Solves with intervals but small ranef
anova(CEX_mod21.103nlme,CEX_mod41.103nlme)
#CEX_mod 4 better
CEX_mod51.103nlme<-update(CEX_mod21.103nlme,random=(R~1),control=iter)
intervals(CEX_mod51.103nlme)
#Solves with intervals
anova(CEX_mod41.103nlme,CEX_mod51.103nlme)
#CEX_mod 5
anova(CEX_mod51.103nlme,CEX_mod11.103nls)
#NLME fit better!
CEX_mod61.103nlme<-update(CEX_mod21.103nlme,random=(C~1),control=iter)
anova(CEX_mod61.103nlme,CEX_mod51.103nlme)
#Mod 5 better! only on R!
summary(CEX_mod51.103nlme)

#########C2 exponential########
##For 34
CEX2_mod11.34nls <- nls(log2game ~ CEx2(day,C,R),data=Data1_34,start=c(C=2.2,R=0.92),trace=T)
#Solves to: 2.22, 0.92
CEX2_mod11.34lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data1_34,start=c(C=2.5,R=0.95))
#Solved for all!
plot(intervals(CEX2_mod11.34lis))
CEX2_mod11.34nlme<-nlme(CEX2_mod11.34lis,random=C+R~1,verbose=T)
intervals(CEX2_mod11.34nlme)
#Solve with intervals
CEX2_mod21.34nlme<-nlme(CEX2_mod11.34lis,random=(C~1),verbose=T)
intervals(CEX2_mod21.34nlme)
#Solves with intervals
anova(CEX2_mod21.34nlme,CEX2_mod11.34nlme)
#CEX2_mod 1 with both ranef better
CEX2_mod31.34nlme<-nlme(CEX2_mod11.34lis,random=pdDiag(C+R~1),verbose=T)
anova(CEX2_mod31.34nlme,CEX2_mod11.34nlme)
#CEX2_mod 1 with both ranef better
CEX2_mod41.34nlme<-nlme(CEX2_mod11.34lis,random=(R~1),verbose=T)
anova(CEX2_mod41.34nlme,CEX2_mod11.34nlme)
#CEX2_mod 1 with both ranef better
summary(CEX2_mod11.34nlme)
```

```
##For 47
CEX2_mod11.47nls <- nls(log2game ~ CEx2(day,C,R),data=Data1_47,start=c(C=2.2,R=0.92),trace=T)
#Solves to: 2.09, 0.93
CEX2_mod11.47lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data1_47,start=c(C=2.3,R=0.95))
#Solved for all!
plot(intervals(CEX2_mod11.47lis))
CEX2_mod11.47nlme<-nlme(CEX2_mod11.47lis,random=C+R~1,verbose=T)
intervals(CEX2_mod11.47nlme)
#Solves with intervals
CEX2_mod21.47nlme<-nlme(CEX2_mod11.47lis,random=pdDiag(C+R~1),verbose=T,control=iter)
intervals(CEX2_mod21.47nlme)
#Solves with intervals
anova(CEX2_mod21.47nlme,CEX2_mod11.47nlme)
#CEX2_mod1 better with unstructured
CEX2_mod31.47nlme<-nlme(CEX2_mod11.47lis,random=(C~1),control=iter)
intervals(CEX2_mod31.47nlme)
#Solves with intervals
anova(CEX2_mod31.47nlme,CEX2_mod11.47nlme)
#CEX2_mod1 better with unstructured
CEX2_mod41.47nlme<-nlme(CEX2_mod11.47lis,random=(R~1),control=iter)
intervals(CEX2_mod41.47nlme)
#Solves with intervals
anova(CEX2_mod41.47nlme,CEX2_mod11.47nlme)
#CEX2_mod1 better with unstructured
summary(CEX2_mod11.47nlme)

##For 56
CEX2_mod11.56nls <- nls(log2game ~ CEx2(day,C,R),data=Data1_56,start=c(C=2.2,R=0.92),trace=T)
#Solves to: 1.96, 0.93
CEX2_mod11.56lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data1_56,start=c(C=2.2,R=0.95))
#Solved for all!
plot(intervals(CEX2_mod11.56lis))
CEX2_mod11.56nlme<-nlme(CEX2_mod11.56lis,random=C+R~1,verbose=T)
intervals(CEX2_mod11.56nlme)
#Solves with intervals!
CEX2_mod21.56nlme<-nlme(CEX2_mod11.56lis,random=pdDiag(C+R~1),verbose=T)
intervals(CEX2_mod21.56nlme)
#Solves with intervals
anova(CEX2_mod21.56nlme,CEX2_mod11.56nlme)
#CEX2_mod1 better with unstructured
CEX2_mod31.56nlme<-nlme(CEX2_mod11.56lis,random=(C~1))
intervals(CEX2_mod31.56nlme)
#Solves with intervals
anova(CEX2_mod31.56nlme,CEX2_mod11.56nlme)
#CEX2_mod1 better with unstructured
CEX2_mod41.56nlme<-nlme(CEX2_mod11.56lis,random=(R~1))
intervals(CEX2_mod41.56nlme)
#Solves with intervals
anova(CEX2_mod41.56nlme,CEX2_mod11.56nlme)
#CEX2_mod1 better with unstructured
summary(CEX2_mod11.56nlme)

##For 74
CEX2_mod11.74nls <- nls(log2game ~ CEx2(day,C,R),data=Data1_74,start=c(C=2.2,R=0.93),trace=T)
#Solves to: 1.99, 0.92
CEX2_mod11.74lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data1_74,start=c(C=2.2,R=0.95))
#Solved for all!
```

```
plot(intervals(CEX2_mod11.74lis))
CEX2_mod11.74nlme<-nlme(CEX2_mod11.74lis,random=C+R~1,verbose=T)
intervals(CEX2_mod11.74nlme)
#Solves with intervals
CEX2_mod21.74nlme<-nlme(CEX2_mod11.74lis,random=pdDiag(C+R~1),verbose=T)
intervals(CEX2_mod21.74nlme)
#Solves with intervals
anova(CEX2_mod21.74nlme,CEX2_mod11.74nlme)
#CEX2_mod1 better with unstructured
CEX2_mod31.74nlme<-nlme(CEX2_mod11.74lis,random=(C~1))
intervals(CEX2_mod31.74nlme)
#Solves with intervals
anova(CEX2_mod31.74nlme,CEX2_mod11.74nlme)
#CEX2_mod1 better with unstructured
CEX2_mod41.74nlme<-nlme(CEX2_mod11.74lis,random=(R~1))
intervals(CEX2_mod41.74nlme)
#Solves with intervals
anova(CEX2_mod41.74nlme,CEX2_mod11.74nlme)
#CEX2_mod1 better with unstructured
summary(CEX2_mod11.74nlme)

##For 88
CEX2_mod11.88nls <- nls(log2game ~ CEx2(day,C,R),data=Data1_88,start=c(C=2.2,R=0.93),trace=T)
#Solves to: 1.91, 0.93
CEX2_mod11.88lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data1_88,start=c(C=2,R=0.95))
#Solved for all!
plot(intervals(CEX2_mod11.88lis))
CEX2_mod11.88nlme<-nlme(CEX2_mod11.88lis,random=C+R~1,verbose=T)
intervals(CEX2_mod11.88nlme)
#Solves with intervals
CEX2_mod21.88nlme<-nlme(CEX2_mod11.88lis,random=pdDiag(C+R~1),verbose=T)
intervals(CEX2_mod21.88nlme)
#Solves with intervals
anova(CEX2_mod21.88nlme,CEX2_mod11.88nlme)
#CEX2_mod1 better with unstructured
CEX2_mod31.88nlme<-nlme(CEX2_mod11.88lis,random=(C~1))
intervals(CEX2_mod31.88nlme)
#Solves with intervals
anova(CEX2_mod31.88nlme,CEX2_mod11.88nlme)
#CEX2_mod1 better with unstructured
CEX2_mod41.88nlme<-nlme(CEX2_mod11.88lis,random=(C~1))
intervals(CEX2_mod41.88nlme)
#Solves with intervals
anova(CEX2_mod41.88nlme,CEX2_mod11.88nlme)
#CEX2_mod1 better with unstructured
summary(CEX2_mod11.88nlme)

##For 103
CEX2_mod11.103nls <- nls(log2game ~ CEx2(day,C,R),data=Data1_103,start=c(C=2.2,R=0.93),trace=T)
#Solves to: 1.8, 0.93
CEX2_mod11.103lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data1_103,start=c(C=2,R=0.95))
#Solved for all!
plot(intervals(CEX2_mod11.103lis))
CEX2_mod11.103nlme<-nlme(CEX2_mod11.103lis,random=C+R~1,verbose=T)
intervals(CEX2_mod11.103nlme)
#Solves with intervals
CEX2_mod21.103nlme<-nlme(CEX2_mod11.103lis,random=pdDiag(C+R~1),verbose=T)
```

```
intervals(CEX2_mod21.103nlme)
#Solves with intervals
anova(CEX2_mod21.103nlme,CEX2_mod11.103nlme)
#CEX2_mod1 better with unstructured
CEX2_mod31.103nlme<-nlme(CEX2_mod11.103lis,random=(C~1))
intervals(CEX2_mod31.103nlme)
#Solves with intervals
anova(CEX2_mod31.103nlme,CEX2_mod11.103nlme)
#CEX2_mod1 better with unstructured
CEX2_mod41.103nlme<-nlme(CEX2_mod11.103lis,random=(R~1))
intervals(CEX2_mod41.103nlme)
#Solves with intervals
anova(CEX2_mod41.103nlme,CEX2_mod11.103nlme)
#CEX2_mod1 better with unstructured

summary(CEX2_mod11.103nlme)

#########Fourier#########

##For 56
F_mod11.56nls <- nls(log2game ~ Fourier(day, A, B,E,W),data=Data1_56,start=c(A=7,B=4,E=7,W=42),trace=T,nls.
control(maxiter=150))
#Solves to: 6.54, 3.84, 6.41, 38.78
F_mod11.56lis <- nlsList(log2game ~ Fourier(day, A, B,E,W),data=Data1_56,start=c(A=7,B=4,E=7,W=42))
#Solved for 13
plot(intervals(F_mod11.56lis))
F_mod11.56nlme<-nlme(F_mod11.56lis,random=A+B+E+W~1,verbose=T)
#Cannot solve, singularity in backsolve
F_mod21.56nlme<-nlme(F_mod11.56lis,random=pdDiag(A+B+E+W~1),verbose=T)
intervals(F_mod21.56nlme)
#Solves but no intervals, all ranef small
F_mod31.56nlme<-update(F_mod21.56nlme,random=(A~1),verbose=T)
intervals(F_mod31.56nlme)
#Solves with intervals!
anova(F_mod31.56nlme,F_mod21.56nlme)
#F_mod 3 better
F_mod41.56nlme<-nlme(F_mod11.56lis,random=pdDiag(A+B+E~1),control=iter)
intervals(F_mod41.56nlme)
#Solves with intervals but shows huge imprecision in ranefs for B E
anova(F_mod31.56nlme,F_mod41.56nlme)
#Mod 3 better
anova(F_mod31.56nlme,F_mod11.56nls)
#NLS fit better

##For 103
F_mod11.103nls <- nls(log2game ~ Fourier(day, A, B,E,W),data=Data1_103,start=c(A=6.3,B=4,E=7,W=42),trace=T,
nls.control(maxiter=200))
#Solves to 5.8, 3.65, 6.28, 38.85
F_mod11.103lis <- nlsList(log2game ~ Fourier(day, A, B,E,W),data=Data1_103,start=c(A=6.3,B=4,E=6.7,W=42))
#Solved for 14
plot(intervals(F_mod11.103lis))
F_mod11.103nlme<-nlme(F_mod11.103lis,random=A+B+E+W~1,verbose=T,control=iter)
#Cannot solve, max iter
F_mod21.103nlme<-nlme(F_mod11.103lis,random=pdDiag(A+B+E+W~1),verbose=T,control=iter)
#Cannot solve, max iter
F_mod31.103nlme<-nlme(F_mod11.103lis,random=(A~1),verbose=T,control=iter)
#Cannot solve, max iter
```

```
F_mod41.103nlme<-nlme(F_mod11.103lis,random=pdDiag(A+B+E~1),verbose=T)
#Cannot solve, max iter

#########Dbl Fourier#########
##For 56
DF_mod11.56nls <- nls(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_56,start=c(A=6.7,B=4.3,E=6.8,W=40,
G=3,P=3),trace=T)
#Solves to: 6.45, 4.15, 6.42, 37.7, 2.78, 2.8
DF_mod11.56lis <- nlsList(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_56,start=c(A=7,B=4.5,E=7,W=40,
G=3,P=3))
#Solves for 1!
DF_mod11.56nlme<-nlme(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_56,fixed=(A+B+E+W+G+P~1),random=(A
+B+E+W+G+P~1),start=c(A=6.7,B=4.3,E=6.8,W=40,G=3,P=3) ,verbose=T)
#error - system computationally singular
DF_mod21.56nlme<-nlme(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_56,fixed=(A+B+E+W+G+P~1),random=pd
Diag(A+B+E+W+G+P~1),start=c(A=6.7,B=4.3,E=6.8,W=40,G=3,P=3),verbose=T)
#Does not solve, step halving
DF_mod31.56nlme<-nlme(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_56,fixed=(A+B+E+W+G+P~1),random=(A
+B+E~1),start=c(A=6.7,B=4.3,E=6.8,W=40,G=3,P=3))
intervals(DF_mod31.56nlme)
#Solves with intervals, much uncertainty around corrs, also ranef for A and B v highly correlated
DF_mod41.56nlme<-update(DF_mod31.56nlme,random=pdDiag(A+B+E~1))
intervals(DF_mod41.56nlme)
#Solves with intervals
anova(DF_mod31.56nlme,DF_mod41.56nlme)
#DF_mod 3 better
DF_mod51.56nlme<-update(DF_mod31.56nlme,random=(A+E~1))
intervals(DF_mod51.56nlme)
#Solves with intervals
anova(DF_mod31.56nlme,DF_mod51.56nlme)
#DF_mod 3 better
DF_mod61.56nlme<-update(DF_mod31.56nlme,random=pdDiag(A+E~1))
intervals(DF_mod61.56nlme)
#Solves with intervals
anova(DF_mod31.56nlme,DF_mod61.56nlme)
#DF_mod 3 better
DF_mod71.56nlme<-update(DF_mod31.56nlme,random=pdBlocked(list(A+B~1,E~1)))
intervals(DF_mod71.56nlme)
#Solves but no intervals
anova(DF_mod31.56nlme,DF_mod71.56nlme)
#DF_mod 7 best i.e. block diagonal structure
anova(DF_mod71.56nlme,DF_mod11.56nls)
#better than nls
summary(DF_mod71.56nlme)

##For 103
DF_mod11.103nls <- nls(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_103,start=c(A=6.5,B=4,E=7,W=40,G=
2.9,P=3.3),trace=T)
#Solves to: 6.02, 3.67, 6.5, 37.4, 2.69, 3.08
DF_mod11.103lis <- nlsList(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_103,start=c(A=6.5,B=4,E=7,W=4
2,G=3,P=3.3))
#Solves for 1
DF_mod11.103nlme<-nlme(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_103,random=(A+B+E+W+G+P~1),fixed=
A+B+E+W+G+P~1, verbose=T,start=c(A=6.5,B=4,E=7,W=40,G=2.9,P=3.3))
intervals(DF_mod11.103nlme)
#Solves but no intervals
DF_mod21.103nlme<-nlme(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data1_103,random=pdDiag(A+B+E+W+G+P~1),
```

```
fixed=A+B+E+W+G+P~1, verbose=T,start=c(A=6.5,B=4,E=7,W=40,G=2.9,P=3.3))
intervals(DF_mod21.103nlme)
#Solves but no intervals, ranef for W,G,P v small
anova(DF_mod21.103nlme, DF_mod11.103nlme)
#DF_mod 1 better
DF_mod31.103nlme<-update(DF_mod21.103nlme,random=pdDiag(A+B+E~1))
intervals(DF_mod31.103nlme)
#Solves with intervals!
anova(DF_mod11.103nlme,DF_mod31.103nlme)
#DF_mod 1 better i.e. unstructured
DF_mod41.103nlme<-update(DF_mod21.103nlme,random=(A+B+E~1))
intervals(DF_mod41.103nlme)
#Solves with intervals!
anova(DF_mod41.103nlme,DF_mod11.103nlme)
#DF_mod 1  better i.e. unstructured
DF_mod51.103nlme<-update(DF_mod21.103nlme,random=pdBlocked(list(A+B~1,E~1)))
intervals(DF_mod51.103nlme)
#Solves but no intervals
anova(DF_mod51.103nlme,DF_mod11.103nlme)
#DF_mod 1 better
summary(DF_mod11.103nlme)

##############################################################################
##### Dataset 5 - day0 zero stays zero, others changed to 8 ##################
##############################################################################
#########PK 1 comp model - para1#########

##For 34
PK1_mod15.34nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_34,start=c(beta0=15,beta1=0.05,be
ta2=0.45),trace=T)
#Solves to 27.96, 0.047, 0.124
PK1_mod15.34lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_34,start=c(beta0=30,beta1=0.0
7,beta2=0.2))
#Solves for 15 subjects
PK1_mod15.34nlme <- nlme(PK1_mod15.34lis,random=(beta0+beta1+beta2~1),verbose=T,control=iter)
intervals(PK1_mod15.34nlme)
#Solves without intervals. Small ranefs
PK1_mod25.34nlme <- nlme(PK1_mod15.34lis,random=pdDiag(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, step halving.
PK1_mod35.34nlme <- nlme(PK1_mod15.34lis,random=pdDiag(beta0+beta1~1),verbose=T,control=iter)
#Cannot solve, max iter.
PK1_mod45.34nlme <- nlme(PK1_mod15.34lis,random=(beta1~1),verbose=T,control=iter)
#Cannot solve, step halving
anova(PK1_mod15.34nlme,PK1_mod15.34nls)
#NLS fit better!

##For 47
PK1_mod15.47nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_47,start=c(beta0=17,beta1=0.05,be
ta2=0.45),trace=T)
#Solves to 22.74, 0.04, 0.133
PK1_mod15.47lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_47,start=c(beta0=25,beta1=0.1
,beta2=0.2))
#Solves for 26 subjects
PK1_mod15.47nlme <- nlme(PK1_mod15.47lis,random=(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, singular precision.
PK1_mod25.47nlme <- nlme(PK1_mod15.47lis,random=pdDiag(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, step halving.
```

```
PK1_mod35.47nlme <- nlme(PK1_mod15.47lis,random=pdDiag(beta0+beta1~1),verbose=T,control=iter)
#Cannot solve, max iter.
PK1_mod45.47nlme <- nlme(PK1_mod15.47lis,random=(beta0~1),verbose=T,control=iter)
#Cannot solve, max iter.
PK1_mod55.47nlme <- nlme(PK1_mod15.47lis,random=(beta1~1),verbose=T,control=iter)
#Cannot solve, step halving.

##For 56
PK1_mod15.56nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_56,start=c(beta0=15,beta1=0.05,be
ta2=0.4),trace=T)
#Solves to 19.676, 0.0355, 0.14
PK1_mod15.56lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_56,start=c(beta0=22,beta1=0.0
5,beta2=0.2))
#Solves for 31 subjects
PK1_mod15.56nlme <- nlme(PK1_mod15.56lis,random=(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, max iter
PK1_mod25.56nlme <- nlme(PK1_mod15.56lis,random=pdDiag(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, step halving.
PK1_mod35.56nlme <- nlme(PK1_mod15.56lis,random=pdDiag(beta0+beta1~1),verbose=T,control=iter)
#Cannot solve, max iter.
PK1_mod45.56nlme <- nlme(PK1_mod15.56lis,random=(beta0~1),verbose=T,control=iter)
intervals(PK1_mod45.56nlme)
#Solves but no intervals
anova(PK1_mod45.56nlme,PK1_mod15.56nls)
#Can solve with only 1 ranef but nls fit better
PK1_mod55.56nlme <- nlme(PK1_mod15.56lis,random=(beta1~1),verbose=T,control=iter)
intervals(PK1_mod25.56nlme)
#Solves but no intervals
anova(PK1_mod55.56nlme,PK1_mod15.56nls)
#Can solve with only 1 ranef but nls fit better

##For 74
PK1_mod15.74nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_74,start=c(beta0=17,beta1=0.05,be
ta2=0.4),trace=T)
#Solves to 21.86, 0.045, 0.13
PK1_mod15.74lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_74,start=c(beta0=25,beta1=0.0
8,beta2=0.2))
#Solves for 32 subjects
PK1_mod15.74nlme <- nlme(PK1_mod15.74lis,random=(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, step halving
PK1_mod25.74nlme <- nlme(PK1_mod15.74lis,random=pdDiag(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, step halving
PK1_mod35.74nlme <- nlme(PK1_mod15.74lis,random=pdDiag(beta0+beta1~1),verbose=T,control=iter)
#Cannot solve, max iter.
PK1_mod45.74nlme <- nlme(PK1_mod15.74lis,random=(beta0~1),verbose=T,control=iter)
intervals(PK1_mod45.74nlme)
#Solves with intervals
PK1_mod55.74nlme <- nlme(PK1_mod15.74lis,random=(beta1~1),verbose=T,control=iter)
#Cannot solve, step halving
anova(PK1_mod45.74nlme, PK1_mod15.74nls)
#NLS fit!

##For 88
PK1_mod15.88nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_88,start=c(beta0=17,beta1=0.05,be
ta2=0.4),trace=T)
#Solves to 19.77, 0.04, 0.14
PK1_mod15.88lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_88,start=c(beta0=23,beta1=0.0
```

```
8,beta2=0.2))
#Solves for 44 subjects
PK1_mod15.88nlme <- nlme(PK1_mod15.88lis,random=(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, singular precision.
PK1_mod25.88nlme <- nlme(PK1_mod15.88lis,random=pdDiag(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, step halving.
PK1_mod35.88nlme <- nlme(PK1_mod15.88lis,random=pdDiag(beta0+beta1~1),verbose=T,control=iter)
#Cannot solve, max iter.
PK1_mod45.88nlme <- nlme(PK1_mod15.88lis,random=(beta0~1),verbose=T,control=iter)
#Cannot solve, max iter.
PK1_mod55.88nlme <- nlme(PK1_mod15.88lis,random=(beta1~1),verbose=T,control=iter)
#Cannot solve, step halving.

##For 103
PK1_mod15.103nls <- nls(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_103,start=c(beta0=17,beta1=0.05,
beta2=0.4),trace=T)
#Solves to 16.69, 0.035, 0.146
PK1_mod15.103lis <- nlsList(log2game ~ PK1.1(day,beta0,beta1,beta2),data=Data5_103,start=c(beta0=20,beta1=0
.1,beta2=0.2))
#Solves for 56 subjects
PK1_mod15.103nlme <- nlme(PK1_mod15.103lis,random=(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, singular precision.
PK1_mod25.103nlme <- nlme(PK1_mod15.103lis,random=pdDiag(beta0+beta1+beta2~1),verbose=T,control=iter)
#Cannot solve, step halving.
PK1_mod35.103nlme <- nlme(PK1_mod15.103lis,random=pdDiag(beta0+beta1~1),verbose=T,control=iter)
#Cannot solve, max iter.
PK1_mod45.103nlme <- nlme(PK1_mod15.103lis,random=(beta0~1),verbose=T,control=iter)
intervals(PK1_mod45.103nlme)
#Solves without intervals!
PK1_mod55.103nlme <- nlme(PK1_mod15.103lis,random=(beta1~1),verbose=T,control=iter)
#Cannot solve, step halving.
anova(PK1_mod45.103nlme, PK1_mod15.103nls)
#nls fit better!

#########Biexponential#########
##For 34
BI_mod15.34nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_34,trace=T)
#Solves to: -26.9, -2.05, 26.73, -3.08
BI_mod15.34lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_34)
#Solves for 15 subjects
BI_mod15.34nlme <- nlme(BI_mod15.34lis,random=(A1+lrc1+A2+lrc2~1),control=iter)
#Cannot solve, max iter and singular precision
BI_mod25.34nlme <- nlme(BI_mod15.34lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
intervals(BI_mod25.34nlme)
#Solves but no intervals, all ranef v small, high corr's!
BI_mod35.34nlme <- nlme(BI_mod15.34lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
intervals(BI_mod35.34nlme)
#Solves but no intervals, ranef v small
anova(BI_mod35.34nlme, BI_mod15.34nls)
#NLS fit better
summary(BI_mod35.34nlme)

##For 47
BI_mod15.47nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_47,trace=T)
#Solves to: -22.21, -1.99, 22.04, -3.24
BI_mod15.47lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_47)
#Solves for 26 subjects
```

```
plot(intervals(BI_mod15.47lis))
BI_mod15.47nlme <- nlme(BI_mod15.47lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, singular precision
BI_mod25.47nlme <- nlme(BI_mod15.47lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving....
BI_mod35.47nlme <- nlme(BI_mod15.47lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve
BI_mod45.47nlme <- nlme(BI_mod15.47lis,random=(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve

##For 56
BI_mod15.56nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_56,trace=T)
#Solves to: -19.44, -1.94, 19.31, -3.35
BI_mod15.56lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_56)
#Solves for 30 subjects
BI_mod15.56nlme <- nlme(BI_mod15.56lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, singularity in backsolve
BI_mod25.56nlme <- nlme(BI_mod15.56lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving
BI_mod35.56nlme <- nlme(BI_mod15.56lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving
BI_mod45.56nlme <- nlme(BI_mod15.56lis,random=(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving

##For 74
BI_mod15.74nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_74,trace=T)
#Solves to: -21.36,-2.01, 21.23, -3.12
BI_mod15.74lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_74)
#Solves for 34 subjects
BI_mod15.74nlme <- nlme(BI_mod15.74lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving
BI_mod25.74nlme <- nlme(BI_mod15.74lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving.
BI_mod35.74nlme <- nlme(BI_mod15.74lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving.
BI_mod45.74nlme <- nlme(BI_mod15.74lis,random=(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving.

##For 88
BI_mod15.88nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_88,trace=T)
#Solves to -19.42, -1.97, 19.29, -3.23
BI_mod15.88lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_88)
#Solves for 46 subjects
BI_mod15.88nlme <- nlme(BI_mod15.88lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, singular precision
BI_mod25.88nlme <- nlme(BI_mod15.88lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
intervals(BI_mod25.88nlme)
#Solves but no intervals, all ranef v small
BI_mod35.88nlme <- nlme(BI_mod15.88lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
intervals(BI_mod35.88nlme)
#Solves without intervals!
anova(BI_mod35.88nlme,BI_mod25.88nlme)
#BI_mod 3 better
anova(BI_mod35.88nlme,BI_mod15.88nls)
#NLS fit better

##For 103
```

```
BI_mod15.103nls <- nls(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_103,trace=T)
#Solves to: -16.56, -1.9, 16.46, -3.37
BI_mod15.103lis <- nlsList(log2game ~ SSbiexp(day,A1, lrc1, A2, lrc2),data=Data5_103)
#Solves for 54 subjects
BI_mod15.103nlme <- nlme(BI_mod15.103lis,random=(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving.
BI_mod25.103nlme <- nlme(BI_mod15.103lis,random=pdDiag(A1+lrc1+A2+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving.
BI_mod35.103nlme <- nlme(BI_mod15.103lis,random=pdDiag(lrc1+lrc2~1),verbose=T,control=iter)
#Cannot solve, step halving

########Triexponential########
##For 34
PK2_mod15.34nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data5_34,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! singular gradient

##For 47
PK2_mod15.47nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data5_47,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! step factor

##For 56
PK2_mod15.56nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data5_56,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! singular gradient

##For 74
PK2_mod15.74nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data5_74,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve! singular gradient

##For 88
PK2_mod15.88nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data5_88,start=c(A1=13,beta1=0
.03,A2=-26,beta2=0.2,A3=13,beta3=0.1),trace=T)
#Does not solve!  singular gradient

##For 103
PK2_mod15.103nls <- nls(log2game ~ PK2.1(day,A1,beta1,A2,beta2,A3,beta3),data=Data5_103,start=c(A1=15,beta1
=0.03,A2=-30,beta2=0.2,A3=15,beta3=0.1),trace=T)
#Does not solve! step factor

#########C exponential#########
##For 34
CEX_mod15.34nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data5_34,start=c(A=1.5,B=-1.6,C=2.2,R=0.93),trace
=T)
#Solves to virtually exact same estimates as Genstat: 1.48, -1.63, 2.05, 0.92
CEX_mod15.34lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data5_34,start=c(A=1.5,B=-1.6,C=2.2,R=0.93))
#Solved for all!!
plot(intervals(CEX_mod15.34lis))

CEX_mod15.34nlme<-nlme(CEX_mod15.34lis,random=(A+B+C+R~1),verbose=T,control=iter)
#Cannot solve, maxiter
CEX_mod25.34nlme<-nlme(CEX_mod15.34lis,random=pdDiag(A+B+C+R~1),verbose=T,control=iter)
intervals(CEX_mod25.34nlme)
#Solves! But no intervals. A, B not sig and highly corr
CEX_mod35.34nlme<-update(CEX_mod25.34nlme,random=pdDiag(B+C+R~1),control=iter)
```

```
intervals(CEX_mod35.34nlme)
#Solves! But no intervals. A, B not sig and highly corr
anova(CEX_mod25.34nlme,CEX_mod35.34nlme)
#CEX_mod 3 better
CEX_mod45.34nlme<-update(CEX_mod25.34nlme,random=(C+R~1),control=iter)
intervals(CEX_mod45.34nlme)
#Solves with intervals! A, B still not sig
anova(CEX_mod35.34nlme,CEX_mod45.34nlme)
#CEX_mod 4 better!
CEX_mod55.34nlme<-update(CEX_mod25.34nlme,random=(C~1),control=iter)
intervals(CEX_mod55.34nlme)
#Solves with intervals! A, B still not sig
anova(CEX_mod55.34nlme,CEX_mod45.34nlme)
#CEX_mod 4 better!
CEX_mod65.34nlme<-update(CEX_mod25.34nlme,random=(R~1),control=iter)
intervals(CEX_mod65.34nlme)
#Solves with intervals
anova(CEX_mod65.34nlme,CEX_mod45.34nlme)
#Mod 4 best!
summary(CEX_mod45.34nlme)

##For 47
CEX_mod15.47nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data5_47,start=c(A=2.1,B=-2.2,C=2,R=0.93),trace=T)
#Solves to: 2.1, -2.26, 1.92, 0.92
CEX_mod15.47lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data5_47,start=c(A=2.2,B=-2.1,C=2.1,R=0.93))
#Solved for all but one!!
plot(intervals(CEX_mod15.47lis))
CEX_mod15.47nlme<-nlme(CEX_mod15.47lis,random=A+B+C+R~1,control=iter)
#not solving, maxiter
CEX_mod25.47nlme<-nlme(CEX_mod15.47lis,random=pdDiag(A+B+C+R~1),verbose=T,control=iter)
intervals(CEX_mod25.47nlme)
#Solves but no intervals, A B not sig and ranef small
CEX_mod35.47nlme<-update(CEX_mod25.47nlme,random=(C+R~1),control=iter)
intervals(CEX_mod35.47nlme)
#Solves with intervals but A, B nonsig plus highly correlated
anova(CEX_mod35.47nlme,CEX_mod25.47nlme)
#CEX_mod 3 better
CEX_mod45.47nlme<-update(CEX_mod25.47nlme,random=pdDiag(C+R~1),control=iter)
intervals(CEX_mod45.47nlme)
anova(CEX_mod35.47nlme,CEX_mod45.47nlme)
#Solves with intervals but CEX_mod 3 better!
CEX_mod55.47nlme<-nlme(CEX_mod15.47lis,random=(C~1))
intervals(CEX_mod55.47nlme)
anova(CEX_mod55.47nlme,CEX_mod35.47nlme)
#Solves with intervals but CEX_mod 3 better!
CEX_mod65.47nlme<-nlme(CEX_mod15.47lis,random=(R~1))
intervals(CEX_mod65.47nlme)
anova(CEX_mod65.47nlme,CEX_mod35.47nlme)
#Solves with intervals but CEX_mod 3 better!
summary(CEX_mod35.47nlme)

##For 56
CEX_mod15.56nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data5_56,start=c(A=2.5,B=-2.6,C=1.8,R=0.93),trace
=T)
#Solves to: 2.52, -2.63, 1.81, 0.92
CEX_mod15.56lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data5_56,start=c(A=2.5,B=-2.6,C=1.8,R=0.93))
#Solved for all but 2!!
```

```
plot(intervals(CEX_mod15.56lis))
CEX_mod15.56nlme<-nlme(CEX_mod15.56lis,random=A+B+C+R^1,verbose=T,control=iter)
#not solving: singular precision
CEX_mod25.56nlme<-nlme(CEX_mod15.56lis,random=pdDiag(A+B+C+R^1),verbose=T,control=iter)
intervals(CEX_mod25.56nlme)
#Solves without intervals, A and B not sig & small ranef
CEX_mod35.56nlme<-nlme(CEX_mod15.56lis,random=(C+R^1),control=iter)
#cannot solve, maxiter
CEX_mod45.56nlme<-nlme(CEX_mod15.56lis,random=pdDiag(C+R^1),control=iter)
intervals(CEX_mod45.56nlme)
#Solves but no intervals. A, B non sig
anova(CEX_mod25.56nlme,CEX_mod45.56nlme)
#CEX_mod 4 better
CEX_mod55.56nlme<-nlme(CEX_mod15.56lis,random=(C^1),control=iter)
intervals(CEX_mod55.56nlme)
#Solves with intervals
anova(CEX_mod55.56nlme,CEX_mod45.56nlme)
#CEX_mod 5 better!
CEX_mod65.56nlme<-nlme(CEX_mod15.56lis,random=(R^1),control=iter)
intervals(CEX_mod65.56nlme)
#Solves without intervals
anova(CEX_mod65.56nlme,CEX_mod55.56nlme)
#CEX_mod 5 better!
summary(CEX_mod55.56nlme)


##For 74
CEX_mod15.74nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data5_74,start=c(A=1.6,B=-1.7,C=1.8,R=0.93),trace
=T)
#Solves to: 1.62, -1.75, 1.74, 0.92
CEX_mod15.74lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data5_74,start=c(A=1.6,B=-1.7,C=1.8,R=0.93))
#Solved for all but 6!!
plot(intervals(CEX_mod15.74lis))
CEX_mod15.74nlme<-nlme(CEX_mod15.74lis,random=A+B+C+R^1,verbose=T,control=iter)
#not solving, maxiter
CEX_mod25.74nlme<-nlme(CEX_mod15.74lis,random=pdDiag(A+B+C+R^1),verbose=T,control=iter)
intervals(CEX_mod25.74nlme)
#Solves with intervals, A and B not sig and huge imprecision
CEX_mod35.74nlme<-update(CEX_mod25.74nlme,random=(C+R^1),control=iter)
intervals(CEX_mod35.74nlme)
#Solves with intervals, A and B not sig
anova(CEX_mod35.74nlme,CEX_mod25.74nlme)
#CEX_mod 3 better!
CEX_mod45.74nlme<-update(CEX_mod25.74nlme,random=pdDiag(C+R^1),control=iter)
intervals(CEX_mod45.74nlme)
anova(CEX_mod45.74nlme,CEX_mod35.74nlme)
#Solves with intervals but unstruc CEX_mod 3 better
CEX_mod55.74nlme<-update(CEX_mod25.74nlme,random=(C^1),control=iter)
intervals(CEX_mod55.74nlme)
anova(CEX_mod55.74nlme,CEX_mod35.74nlme)
#Solves with intervals but CEX_mod 3 better
summary(CEX_mod35.74nlme)

##For 88
CEX_mod15.88nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data5_88,start=c(A=2,B=-2,C=1.8,R=0.93),trace=T)
#Solves to : 1.99, -2.12, 1.72, 0.92
CEX_mod15.88lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data5_88,start=c(A=2,B=-2,C=1.8,R=0.93))
```

```
#Solved for all but 8!!
plot(intervals(CEX_mod15.88lis))
CEX_mod15.88nlme<-nlme(CEX_mod15.88lis,random=A+B+C+R~1,verbose=T,control=iter)
#Cannot solve
CEX_mod25.88nlme<-nlme(CEX_mod15.88lis,random=pdDiag(A+B+C+R~1),verbose=T,control=iter)
intervals(CEX_mod25.88nlme)
#Solves without intervals, A and B not sig and small ranef
CEX_mod35.88nlme<-update(CEX_mod25.88nlme,random=(C+R~1)intervals(CEX_mod25.88nlme))
intervals(CEX_mod35.88nlme)
#Solves with intervals, A B not sig
anova(CEX_mod35.88nlme,CEX_mod25.88nlme)
#CEX_mod 3 (unstructured) better
CEX_mod45.88nlme<-update(CEX_mod25.88nlme,random=pdDiag(C+R~1),control=iter)
intervals(CEX_mod45.88nlme)
anova(CEX_mod35.88nlme,CEX_mod45.88nlme)
#Solves with intervals but CEX_mod 3 (unstructured) better
CEX_mod55.88nlme<-update(CEX_mod25.88nlme,random=(C~1),control=iter)
intervals(CEX_mod55.88nlme)
anova(CEX_mod55.88nlme,CEX_mod35.88nlme)
#Solves with intervals but CEX_mod 3 (unstructured) better
CEX_mod65.88nlme<-update(CEX_mod25.88nlme,random=(R~1),control=iter)
intervals(CEX_mod65.88nlme)
anova(CEX_mod65.88nlme,CEX_mod35.88nlme)
#Solves with intervals but CEX_mod 3 (unstructured) better
summary(CEX_mod35.88nlme)

##For 103
CEX_mod15.103nls <- nls(log2game ~ CEx(day, A, B,C,R),data=Data5_103,start=c(A=2.5,B=-2.5,C=1.6,R=0.92),tra
ce=T)
#Solves to: 2.46, -2.56, 1.62, 0.91
CEX_mod15.103lis <- nlsList(log2game ~ CEx(day, A, B,C,R),data=Data5_103,start=c(A=2.6,B=-2.2,C=1.9,R=0.92))
#Solved for all but 10!!
plot(intervals(CEX_mod15.103lis))
CEX_mod15.103nlme<-nlme(CEX_mod15.103lis,random=A+B+C+R~1,verbose=T,control=iter)
#not solving, maxiter
CEX_mod25.103nlme<-nlme(CEX_mod15.103lis,random=pdDiag(A+B+C+R~1),verbose=T,control=iter)
intervals(CEX_mod25.103nlme)
#Solves with intervals, A and B not sig and huge imprecision, also for R
CEX_mod35.103nlme<-update(CEX_mod25.103nlme,random=(C+R~1),control=iter)
#Cannot solve, maxiter
CEX_mod45.103nlme<-update(CEX_mod25.103nlme,random=pdDiag(C+R~1),control=iter)
intervals(CEX_mod45.103nlme)
#Solves with intervals, A and B not sig and huge imprecision for R
anova(CEX_mod45.103nlme,CEX_mod25.103nlme)
#CEX_mod 4 better
CEX_mod55.103nlme<-update(CEX_mod25.103nlme,random=(C~1),control=iter)
intervals(CEX_mod55.103nlme)
#Solves with intervals
anova(CEX_mod55.103nlme,CEX_mod45.103nlme)
#CEX_mod 5 better i.e. with only a ranef for C!
CEX_mod65.103nlme<-update(CEX_mod25.103nlme,random=(R~1),control=iter)
intervals(CEX_mod65.103nlme)
#Solves without intervals
anova(CEX_mod55.103nlme,CEX_mod65.103nlme)
#CEX_mod 5 better i.e. with only a ranef for C!
summary(CEX_mod55.103nlme)
```

```
########Cex2########
##For 34
CEX2_mod15.34nls <- nls(log2game ~ CEx2(day,C,R),data=Data5_34,start=c(C=2.1,R=0.93),trace=T)
#Solves 2.03, 0.93
CEX2_mod15.34lis <- nlsList(log2game ~ CEx2(day, C,R),data=Data5_34,start=c(C=2.1,R=0.93))
#Solved for all!!
plot(intervals(CEX2_mod15.34lis))
CEX2_mod15.34nlme<-nlme(CEX2_mod15.34lis,random=C+R~1,verbose=T,control=iter)
intervals(CEX2_mod15.34nlme)
#Solves with intervals
CEX2_mod25.34nlme<-nlme(CEX2_mod15.34lis,random=C~1,verbose=T,control=iter)
intervals(CEX2_mod25.34nlme)
anova(CEX2_mod25.34nlme,CEX2_mod15.34nlme)
#Solves with intervals but CEX2_model with both ranefs better!
CEX2_mod35.34nlme<-nlme(CEX2_mod15.34lis,random=pdDiag(C+R~1),verbose=T,control=iter)
intervals(CEX2_mod35.34nlme)
anova(CEX2_mod35.34nlme,CEX2_mod15.34nlme)
#CEX2_mod 1 better than diag structure
CEX2_mod45.34nlme<-nlme(CEX2_mod15.34lis,random=(R~1),verbose=T,control=iter)
intervals(CEX2_mod45.34nlme)
anova(CEX2_mod45.34nlme,CEX2_mod15.34nlme)
#CEX2_mod 1 better
summary(CEX2_mod15.34nlme)

##For 47
CEX2_mod15.47nls <- nls(log2game ~ CEx2(day,C,R),data=Data5_47,start=c(C=2,R=0.93),trace=T)
#Solves to: 1.92, 0.93
CEX2_mod15.47lis <- nlsList(log2game ~ CEx2(day, C,R),data=Data5_47,start=c(C=2,R=0.93))
#Solved for all!!
plot(intervals(CEX2_mod15.47lis))
CEX2_mod15.47nlme<-nlme(CEX2_mod15.47lis,random=C+R~1,verbose=T)
intervals(CEX2_mod15.47nlme)
#Solves with intervals
CEX2_mod25.47nlme<-nlme(CEX2_mod15.47lis,random=C~1,verbose=T)
intervals(CEX2_mod25.47nlme)
anova(CEX2_mod25.47nlme,CEX2_mod15.47nlme)
#Solves with intervals but CEX2_model with both ranefs better!
CEX2_mod35.47nlme<-nlme(CEX2_mod15.47lis,random=pdDiag(C+R~1),verbose=T)
intervals(CEX2_mod35.47nlme)
anova(CEX2_mod35.47nlme,CEX2_mod15.47nlme)
#CEX2_mod 1 better than diag structure
CEX2_mod45.47nlme<-nlme(CEX2_mod15.47lis,random=(R~1),verbose=T)
intervals(CEX2_mod45.47nlme)
anova(CEX2_mod45.47nlme,CEX2_mod15.47nlme)
#CEX2_mod 1 better
summary(CEX2_mod15.47nlme)

##For 56
CEX2_mod15.56nls <- nls(log2game ~ CEx2(day,C,R),data=Data5_56,start=c(C=1.8,R=0.93),trace=T)
#Solves to: 1.82, 0.93
CEX2_mod15.56lis <- nlsList(log2game ~ CEx2(day, C,R),data=Data5_56,start=c(C=1.8,R=0.93))
#Solved for all!!
plot(intervals(CEX2_mod15.56lis))
CEX2_mod15.56nlme<-nlme(CEX2_mod15.56lis,random=C+R~1,verbose=T)
intervals(CEX2_mod15.56nlme)
#Solves with intervals
CEX2_mod25.56nlme<-nlme(CEX2_mod15.56lis,random=C~1,verbose=T)
```

```
intervals(CEX2_mod25.56nlme)
anova(CEX2_mod25.56nlme,CEX2_mod15.56nlme)
#Solves with intervals but CEX2_model with both ranefs better!
CEX2_mod35.56nlme<-nlme(CEX2_mod15.56lis,random=pdDiag(C+R~1),verbose=T)
intervals(CEX2_mod35.56nlme)
anova(CEX2_mod35.56nlme,CEX2_mod15.56nlme)
#CEX2_mod 1 better than diag structure
CEX2_mod45.56nlme<-nlme(CEX2_mod15.56lis,random=(R~1),verbose=T)
intervals(CEX2_mod45.56nlme)
anova(CEX2_mod45.56nlme,CEX2_mod15.56nlme)
#CEX2_mod 1 better
summary(CEX2_mod15.56nlme)

##For 74
CEX2_mod15.74nls <- nls(log2game ~ CEx2(day,C,R),data=Data5_74,start=c(C=1.8,R=0.93),trace=T)
#Solves to: 1.74, 0.926
CEX2_mod15.74lis <- nlsList(log2game ~ CEx2(day, C,R),data=Data5_74,start=c(C=1.8,R=0.93))
#Solved for all!!
plot(intervals(CEX2_mod15.74lis))
CEX2_mod15.74nlme<-nlme(CEX2_mod15.74lis,random=C+R~1,verbose=T)
intervals(CEX2_mod15.74nlme)
#Solves with intervals
CEX2_mod25.74nlme<-nlme(CEX2_mod15.74lis,random=C~1,verbose=T)
intervals(CEX2_mod25.74nlme)
anova(CEX2_mod25.74nlme,CEX2_mod15.74nlme)
#Solves with intervals but CEX2_model with both ranefs better!
CEX2_mod35.74nlme<-nlme(CEX2_mod15.74lis,random=pdDiag(C+R~1),verbose=T)
intervals(CEX2_mod35.74nlme)
anova(CEX2_mod35.74nlme,CEX2_mod15.74nlme)
#CEX2_mod 1 better than diag structure
CEX2_mod45.74nlme<-nlme(CEX2_mod15.74lis,random=(R~1),verbose=T)
intervals(CEX2_mod45.74nlme)
anova(CEX2_mod45.74nlme,CEX2_mod15.74nlme)
#CEX2_mod 1 better
summary(CEX2_mod15.74nlme)

##For 88
CEX2_mod15.88nls <- nls(log2game ~ CEx2(day,C,R),data=Data5_88,start=c(C=1.8,R=0.93),trace=T)
#Solves to: 1.72, 0.928
CEX2_mod15.88lis <- nlsList(log2game ~ CEx2(day, C,R),data=Data5_88,start=c(C=1.8,R=0.93))
#Solved for all!!
plot(intervals(CEX2_mod15.88lis))
CEX2_mod15.88nlme<-nlme(CEX2_mod15.88lis,random=C+R~1,verbose=T)
intervals(CEX2_mod15.88nlme)
#Solves with intervals
CEX2_mod25.88nlme<-nlme(CEX2_mod15.88lis,random=C~1,verbose=T)
intervals(CEX2_mod25.88nlme)
anova(CEX2_mod25.88nlme,CEX2_mod15.88nlme)
#Solves with intervals but CEX2_model with both ranefs better!
CEX2_mod35.88nlme<-nlme(CEX2_mod15.88lis,random=pdDiag(C+R~1),verbose=T)
intervals(CEX2_mod35.88nlme)
anova(CEX2_mod35.88nlme,CEX2_mod15.88nlme)
#CEX2_mod 1 better than diag structure
CEX2_mod45.88nlme<-nlme(CEX2_mod15.88lis,random=(C~1),verbose=T)
intervals(CEX2_mod45.88nlme)
anova(CEX2_mod45.88nlme,CEX2_mod15.88nlme)
#CEX2_mod 1 better than diag structure
```

```
summary(CEX2_mod15.88nlme)

##For 103
CEX2_mod15.103nls <- nls(log2game ~ CEx2(day,C,R),data=Data5_103,start=c(C=1.7,R=0.93),trace=T)
#Solves to: 1.64, 0.93
CEX2_mod15.103lis <- nlsList(log2game ~ CEx2(day, C,R),data=Data5_103,start=c(C=1.7,R=0.93))
#Solved for all!!
plot(intervals(CEX2_mod15.103lis))
CEX2_mod15.103nlme<-nlme(CEX2_mod15.103lis,random=C+R~1,verbose=T)
intervals(CEX2_mod15.103nlme)
#Solves with intervals
CEX2_mod25.103nlme<-nlme(CEX2_mod15.103lis,random=C~1,verbose=T)
intervals(CEX2_mod25.103nlme)
anova(CEX2_mod25.103nlme,CEX2_mod15.103nlme)
#Solves with intervals but CEX2_model with both ranefs better!
CEX2_mod35.103nlme<-nlme(CEX2_mod15.103lis,random=pdDiag(C+R~1),verbose=T)
intervals(CEX2_mod35.103nlme)
anova(CEX2_mod35.103nlme,CEX2_mod15.103nlme)
#CEX2_mod 1 better than diag structure
CEX2_mod45.103nlme<-nlme(CEX2_mod15.103lis,random=(R~1),verbose=T)
intervals(CEX2_mod45.103nlme)
anova(CEX2_mod45.103nlme,CEX2_mod15.103nlme)
#CEX2_mod 1 better than diag structure
summary(CEX2_mod15.103nlme)

#########Fourier#########

##For 56
F_mod15.56nls <- nls(log2game ~ Fourier(day, A, B,E,W),data=Data5_56,start=c(A=6.5,B=4.5,E=7,W=40),trace=T)
#Solves to 6.19, 4, 6.54, 39.68
F_mod15.56lis <- nlsList(log2game ~ Fourier(day, A, B,E,W),data=Data5_56,start=c(A=6.2,B=4,E=6.5,W=40))
#Solved for all but 14
plot(intervals(F_mod15.56lis))
F_mod15.56nlme<-nlme(F_mod15.56lis,random=A+B+E+W~1,verbose=T,control=iter)
intervals(F_mod15.56nlme)
#Solves but no intervals
F_mod25.56nlme<-nlme(F_mod15.56lis,random=pdDiag(A+B+E+W~1),verbose=T)
intervals(F_mod25.56nlme)
#Solves but no intervals, high corr btwn E & W
anova(F_mod15.56nlme,F_mod25.56nlme)
#F_mod1 better
F_mod35.56nlme<-nlme(F_mod15.56lis,random=(A+B+E~1),verbose=T)
#Cannot solve, computationally singular
F_mod45.56nlme<-nlme(F_mod15.56lis,random=pdDiag(A+B+E~1),verbose=T)
intervals(F_mod45.56nlme)
anova(F_mod15.56nlme,F_mod45.56nlme)
#Solves but no intervals, ranef small and F_mod 1 better
anova(F_mod15.56nlme,F_mod15.56nls)
#debatable which is better, LR test says maybe NLME, AIC says NLS
summary(F_mod15.56nlme)

##For 103
F_mod15.103nls <- nls(log2game ~ Fourier(day, A, B,E,W),data=Data5_103,start=c(A=5.4,B=3.7,E=6.5,W=40),trac
e=T,nls.control(maxiter=200))
#Solves to 5.4, 3.67, 6.33, 39.23
F_mod15.103lis <- nlsList(log2game ~ Fourier(day, A, B,E,W),data=Data5_103,start=c(A=6,B=4,E=7,W=42))
#Solved for all but 26
```

```
plot(intervals(F_mod15.103lis))
F_mod15.103nlme<-nlme(F_mod15.103lis,random=A+B+E+W~1,verbose=T,control=iter)
intervals(F_mod15.103nlme)
#Solves but no confidence intervals, high corr btwn E & W
F_mod25.103nlme<-nlme(F_mod15.103lis,random=pdDiag(A+B+E+W~1),verbose=T)
intervals(F_mod25.103nlme)
#Solves but no intervals, ranef small
anova(F_mod25.103nlme,F_mod15.103nlme)
#F_mod 1 better
F_mod35.103nlme<-nlme(F_mod15.103lis,random=(A+B+E~1),verbose=T)
intervals(F_mod35.103nlme)
#Solves but no intervals
anova(F_mod35.103nlme,F_mod15.103nlme)
#F_mod 3 better i.e. without ranef for W, A & B high corr though
F_mod45.103nlme<-nlme(F_mod15.103lis,random=(B+E~1),verbose=T)
intervals(F_mod45.103nlme)
anova(F_mod35.103nlme,F_mod45.103nlme)
#Solves but no intervals, F_mod 3 better
anova(F_mod35.103nlme,F_mod15.103nls)
#F_mod with ranef better
summary(F_mod35.103nlme)

#########Dbl Fourier#########
##For 56
DF_mod15.56nls <- nls(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data5_56,start=c(A=6,B=4.5,E=6.5,W=40,G=
2.5,P=2.8),trace=T)
#Solves to: 6.04, 4.36, 6.52, 38.64, 2.42, 2.79
DF_mod15.56lis <- nlsList(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data5_56,start=c(A=6.5,B=5,E=7,W=42,
G=2.8,P=3))
#Solves for all but 5
DF_mod15.56nlme<-nlme(DF_mod15.56lis,random=(A+B+E+W+G+P~1),verbose=T)
#Solves without intervals, ranefs appear larger
DF_mod25.56nlme<-nlme(DF_mod15.56lis,random=pdDiag(A+B+E+W+G+P~1),verbose=T)
intervals(DF_mod25.56nlme)
#Solves without intervals
anova(DF_mod15.56nlme,DF_mod25.56nlme)
#DF_mod 1 better
DF_mod35.56nlme<-update(DF_mod25.56nlme,random=(A+B+E~1))
intervals(DF_mod35.56nlme)
#Solves but no intervals
anova(DF_mod15.56nlme,DF_mod35.56nlme)
#DF_mod 1 better
DF_mod45.56nlme<-update(DF_mod25.56nlme,random=pdDiag(A+B+E~1))
intervals(DF_mod45.56nlme)
anova(DF_mod45.56nlme,DF_mod15.56nlme)
#Solves with intervals but DF_mod 1 better
DF_mod55.56nlme<-update(DF_mod25.56nlme,random=(B+E~1))
intervals(DF_mod55.56nlme)
anova(DF_mod55.56nlme,DF_mod15.56nlme)
#Solves without intervals but DF_mod 1 better
DF_mod65.56nlme<-update(DF_mod25.56nlme,random=pdDiag(A+B+E+W~1))
intervals(DF_mod65.56nlme)
anova(DF_mod65.56nlme,DF_mod15.56nlme)
#Solves with intervals, DF_mod 1 better
anova(D5DF_mod15.56nlme,DF_mod15.56nls)
#NLME better
summary(DF_mod15.56nlme)
```

```
##For 103
DF_mod15.103nls <- nls(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data5_103,start=c(A=5.5,B=4,E=6.5,W=40,
G=2,P=3),trace=T)
#Solves to: 5.3, 3.94, 6.33, 38.58, 2, 2.77
DF_mod15.103lis <- nlsList(log2game ~ DbFourier(day, A, B,E,W,G,P),data=Data5_103,start=c(A=5.5,B=4,E=7,W=4
0,G=2.2,P=3))
#Solves for all but 16
DF_mod15.103nlme<-nlme(DF_mod15.103lis,random=(A+B+E+W+G+P~1),verbose=T)
intervals(DF_mod15.103nlme)
#Solves without intervals!
DF_mod25.103nlme<-nlme(D5DF_mod15.103lis,random=pdDiag(A+B+E+W+G+P~1),verbose=T)
intervals(DF_mod25.103nlme)
#Solves without intervals, ranef for W,G,P v small
anova(DF_mod15.103nlme,DF_mod25.103nlme)
#DF_mod 1 better
DF_mod35.103nlme<-update(DF_mod25.103nlme,random=(A+B+E~1))
intervals(DF_mod35.103nlme)
anova(DF_mod35.103nlme,DF_mod15.103nlme)
#Solves with intervals! Corr (BE) very uncertain, DF_mod 1 better
DF_mod45.103nlme<-update(DF_mod25.103nlme,random=pdBlocked(list(A+B~1,E~1)))
intervals(DF_mod45.103nlme)
anova(DF_mod45.103nlme,DF_mod15.103nlme)
#Solves but no intervals, DF_mod 1 better
DF_mod55.103nlme<-update(DF_mod25.103nlme,random=pdDiag(B+E~1))
intervals(DF_mod55.103nlme)
anova(DF_mod55.103nlme,DF_mod15.103nlme)
#Solves but no intervals, DF_mod 1 better
DF_mod65.103nlme<-update(DF_mod25.103nlme,random=pdDiag(A+B+E~1))
intervals(DF_mod65.103nlme)
anova(DF_mod65.103nlme,DF_mod15.103nlme)
#Solves with intervals, but DF_mod 1 better
anova(DF_mod15.103nlme,DF_mod15.103nls)
#DF_mod with ranefs better
summary(DF_mod15.103nlme)
```

Exploring the final structure

```
x<-seq(0,70,length=10000)
CEx2 <- function (X,C,R) {
(C*X)*(R^X)}
T1 <- function (X,C) {
(C*X)}
T2 <- function (X,R) {
(R^X)}
y1<-CEx2(x,1,0.93)
y2<-CEx2(x,2,0.93)
y3<-CEx2(x,3,0.93)
y4<-CEx2(x,3,0.4)
y5<-CEx2(x,3,0.7)
y6<-CEx2(x,3,0.9)
T1<-T1(x,3)
T2<-T2(x,0.93)
Co<-CEx2(x,3,0.93)

plot(c(0,70),c(0,100),type='n',xlab='Days', ylab='y')
lines(x,T1,type="l", lwd=2, col=1)
```

```
lines(x,Co,type="l", lty=2,lwd=2, col=1)
par(new=T)
plot(c(0,70),c(0,1),type='n',xlab='Days', ylab='y',axes=F)
lines(x,T2,type="l", lwd=2, col='blue')
axis(side=4,col.axis='blue')
legend(40,0.8,cex=0.85, legend=c('Term1','Term2', 'Combined '),
    lty=c(1,1,2),col=c(1,'blue',1))
title('Exploring the relative contributions of the 2 terms')

par(mfcol=c(1,2))
plot(c(0,70),c(0,15),type='n',xlab='Days', ylab='y')
lines(x,y1,type="l", lwd=2, col=2)
lines(x,y2, type="l", lwd=2, col=3)
lines(x,y3, type="l", lwd=2, col=4)
legend(28,14,cex=0.6, legend=c('C=1 ; R=0.93', 'C=2 ; R=0.93', 'C=3 ; R=0.93 '),
    lty=1,col=c(2,3,4))
title('Exploring the C Parameter')

plot(c(0,70),c(0,12),type='n',xlab='Days', ylab='y')
lines(x,y4,type="l", lwd=2, col=2)
lines(x,y5, type="l", lwd=2, col=3)
lines(x,y6, type="l", lwd=2, col=4)
legend(28,11,cex=0.6, legend=c('C=3 ; R=0.4', 'C=3 ; R=0.7', 'C=3 ; R=0.9 '),
    lty=1,col=c(2,3,4))
title('Exploring the R Parameter')
```

D.2.2 Phase 2 - modeling covariates

```
#######################################################################
########Read in data########
####Dataset 1 - only day0 zeros
Data<-read.csv("Final_Data1.csv",header=T,sep=",")
attach(Data)
Data1<-transform(Data,log2game=log2(gamedens+1),pcat=as.factor(pcat),logpdens=log10(pdens0))
detach(Data)
Data1<-groupedData(log2game ~ day | subjectno, data=Data1)

Data<-read.csv("Final_Data1R.csv",header=T,sep=",")
attach(Data)
Data1R<-transform(Data,log2game=log2(gamedens+1),pcat=as.factor(pcat),logpdens=log10(pdens0))
detach(Data)
Data1R<-groupedData(log2game ~ day | subjectno, data=Data1R)

####Dataset 2 - day0 left as 0 and others changed to 8
Data<-read.csv("Final_Data2.csv",header=T,sep=",")
attach(Data)
Data2<-transform(Data,log2game=log2(gamedens+1),pcat=as.factor(pcat),logpdens=log10(pdens0))
detach(Data)
Data2<-groupedData(log2game ~ day | subjectno, data=Data2)

Data<-read.csv("Final_Data2R.csv",header=T,sep=",")
attach(Data)
Data2R<-transform(Data,log2game=log2(gamedens+1),pcat=as.factor(pcat),logpdens=log10(pdens0))
detach(Data)
Data2R<-groupedData(log2game ~ day | subjectno, data=Data2R)

#######################################################################
```

```
#######Models######
iter<-nlmeControl(maxIter=500)

##############################################################################
#####Dataset 1 - only day0 zeros##############################################
##############################################################################

#No covariates
D1C21.nls <- nls(log2game ~ CEx2(day,C,R),data=Data1,start=c(C=2.2,R=0.93),trace=T)
#Solves to: 1.81, 0.93
D1C21.lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data1,start=c(C=2,R=0.95),na.action=na.pass)
#Solved for all!
D1C21.nlme<-nlme(D1C21.lis,random=C+R~1,verbose=T)
intervals(D1C21.nlme)
summary(D1C21.nlme)
D1C2.ran1 <- ranef(D1C21.nlme,augFrame=T)
plot(D1C2.ran1,form= C ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat,fail,site and logpdens, maybe age
plot(D1C2.ran1,form= R ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#quite a few promising covs! Age, fail, pcat, site,pct

#1st cov: age
D1C22.nlme <- update (D1C21.nlme,fixed=list(C ~ 1, R ~ age),start=c(1.8777,0.9279,0))
summary(D1C22.nlme)
D1C23.nlme <- update (D1C21.nlme,fixed=list(C ~ age, R ~ age),start=c(1.8777,0,0.9279,0))
summary(D1C23.nlme)
anova(D1C23.nlme,D1C22.nlme)
#better with age on both
anova(D1C23.nlme,D1C21.nlme)
anova(D1C23.nlme,Terms=c(2,4))
#joint significance
D1C2.ran2 <- ranef(D1C23.nlme,augFrame=T)
plot(D1C2.ran2,form= C.(Intercept) ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat, site, outcome, and logpdens
plot(D1C2.ran2,form= R.(Intercept) ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#fail, pcat, site, pct

#2nd cov: outcome
D1C24.nlme <- update (D1C21.nlme,fixed=list(C ~ age+fail, R ~ age+fail),start=c(1.8777,0,0,0.9279,0,0),na.a
ction=na.omit)
summary(D1C24.nlme)
#not sig for either
D1C25.nlme <- update (D1C21.nlme,fixed=list(C ~ age, R ~ age+fail),start=c(1.8777,0,0.9279,0,0),na.action=n
a.omit)
summary(D1C25.nlme)
anova(D1C25.nlme,D1C24.nlme)
#mod with only on R better but not sig
Data<-read.csv("Final_Data1_102.csv",header=T,sep=",")
attach(Data)
Data102<-transform(Data,log2game=log2(gamedens+1),pcat=as.factor(pcat),logpdens=log10(pdens0))
detach(Data)
Data102<-groupedData(log2game ~ day | subjectno, data=Data102)
testfail.lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data102,start=c(C=2,R=0.95),na.action=na.pass)
#Solved for all!
testfail1.nlme<-nlme(testfail.lis,random=C+R~1)
testfail2.nlme <- update (testfail1.nlme,fixed=list(C ~ age, R ~ age),start=c(1.8777,0,0.9279,0))
testfail3.nlme <- update (testfail1.nlme,fixed=list(C ~ age, R ~ age+fail),start=c(1.8777,0,0.9279,0,0))
```

```
anova(testfail3.nlme,testfail2.nlme)

#2nd cov: site
D1C26.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site), R ~ age+factor(site)),start=c(1.8777,0,0,
0.9279,0,0))
summary(D1C26.nlme)
#sig for both but affects sig of C.age
D1C27.nlme <- update (D1C21.nlme,fixed=list(C ~ age, R ~ age+factor(site)),start=c(1.8777,0,0.9279,0,0),na.
action=na.omit)
summary(D1C27.nlme)
anova(D1C27.nlme,D1C26.nlme)
#better with site on both
D1C28.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site), R ~ age),start=c(1.8777,0,0,0.9279,0),na.
action=na.omit)
summary(D1C28.nlme)
#site not sig if used only on C
anova(D1C28.nlme,D1C26.nlme)
anova(D1C26.nlme,D1C23.nlme)
#model improved by adding site to both
anova(D1C26.nlme,Terms=c(3,6))
#joint significance
D1C2.ran3 <- ranef(D1C26.nlme,augFrame=T)
plot(D1C2.ran3,form= C.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat and logpdens
plot(D1C2.ran3,form= R.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat, pct

#3rd cov: logpdens
D1C29.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens, R ~ age+factor(site)),start=c(1.
8777,0,0,0,0.9279,0,0))
summary(D1C29.nlme)
D1C210.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens, R ~ age+factor(site)+logpdens),
start=c(1.8777,0,0,0,0.9279,0,0,0))
summary(D1C210.nlme)
#neither sig when on both parameters
anova(D1C210.nlme,D1C29.nlme)
#better with logpdens only on C
D1C211.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site), R ~ age+factor(site)+logpdens),start=c(1
.8777,0,0,0.9279,0,0,0))
summary(D1C211.nlme)
anova(D1C211.nlme,D1C29.nlme)
#mod 9 better i.e. with logpdens on C
anova(D1C29.nlme,D1C26.nlme)
#better with logpdens
D1C2.ran4 <- ranef(D1C29.nlme,augFrame=T)
plot(D1C2.ran4,form= C.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat and maybe pct?
plot(D1C2.ran4,form= R.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat, maybe pct?

#4th cov: pct
D1C212.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens+pct, R ~ age+factor(site)+pct),s
tart=c(1.8777,0,0,0,0,0.9279,0,0,0))
summary(D1C212.nlme)
#not sig for both
D1C213.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens+pct, R ~ age+factor(site)),start
=c(1.8777,0,0,0,0,0.9279,0,0))
```

```
summary(D1C213.nlme)
#not sig on C only
D1C214.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens, R ~ age+factor(site)+pct),start
=c(1.8777,0,0,0,0.9279,0,0,0))
summary(D1C214.nlme)
#not sig on R only
anova(D1C214.nlme,D1C212.nlme)
anova(D1C214.nlme,D1C213.nlme)
#best version with pct on R only but not sig
anova(D1C214.nlme,D1C29.nlme)
#pct does not help! trying pct as an ordered factor
Data1$pct2<-ordered(Data1$pct)
D1C2.ran4 <- ranef(D1C29.nlme,augFrame=T)
plot(D1C2.ran4,form= C.(Intercept) ~ pct2)
#pcat and maybe pct?
plot(D1C2.ran4,form= R.(Intercept) ~ pct2)
#no evidence for pct as ordered factor

#4th cov: pcat
D1C216.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens+factor(pcat), R ~ age+factor(sit
e)+factor(pcat)),start=c(1.8777,0,0,0,0,0,0.9279,0,0,0,0))
summary(D1C216.nlme)
D1C217.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens, R ~ age+factor(site)+factor(pca
t)),start=c(1.8777,0,0,0,0.9279,0,0,0,0))
summary(D1C217.nlme)
anova(D1C217.nlme,D1C216.nlme)
#better with pcat for both parameters
D1C218.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens+factor(pcat), R ~ age+factor(sit
e)),start=c(1.8777,0,0,0,0,0,0.9279,0,0))
summary(D1C217.nlme)
anova(D1C218.nlme,D1C216.nlme)
#better with pcat on both
anova(D1C216.nlme,D1C29.nlme)
#improves model
anova(D1C216.nlme,Terms=c(5))
D1C2.ran5 <- ranef(D1C216.nlme,augFrame=T)
plot(D1C2.ran5,form= C.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#nothing
plot(D1C2.ran5,form= R.(Intercept) ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#fail looks promising but not shown not to be sig

#Dropping age
D1C219.nlme <- update (D1C21.nlme,fixed=list(C ~ factor(site)+logpdens+factor(pcat), R ~ age+factor(site)+f
actor(pcat)),start=c(1.8777,0,0,0,0,0.9279,0,0,0,0))
summary(D1C219.nlme)
anova(D1C219.nlme,D1C216.nlme)
#p of 0.06 on LR test, marginal!

####Trying few other things
#checking interaction with mut and site
D1C21R.lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data1R,start=c(C=2,R=0.95),na.action=na.pass)
#Solved for all!
D1C21R.nlme<-nlme(D1C21R.lis,random=C+R~1,verbose=T)
D1C22R.nlme <- update (D1C21R.nlme,fixed=list(C ~ age+factor(site)+logpdens+factor(pcat), R ~ age+factor(si
te)+factor(pcat)),start=c(1.8777,0,0,0,0,0,0.9279,0,0,0,0))
D1C23R.nlme <- update (D1C21R.nlme,fixed=list(C ~ age+factor(site)*factor(mut_3)+logpdens+factor(pcat), R ~ a
summary(D1C23R.nlme)
```

```
#interaction term not sig
anova(D1C23R.nlme,D1C22R.nlme)
#no improvement
D1C24R.nlme <- update (D1C21R.nlme,fixed=list(C ~ age+factor(site)+factor(mut_3)+logpdens+factor(pcat), R ~ a
anova(D1C24R.nlme,D1C22R.nlme)
#no improvement
D1C25R.nlme <- update (D1C21R.nlme,fixed=list(C ~ age+factor(site)*factor(mut_3)+logpdens+factor(pcat), R ~ a
anova(D1C25R.nlme,D1C22R.nlme)
#no improvement
D1C2R.ran1 <- ranef(D1C22R.nlme,augFrame=T)
plot(D1C2R.ran1, form = ~ site*mut_3)

#Final model - diagnostics
summary(D1C216.nlme)
plot(D1C216.nlme, log2game ~ fitted(.), abline=c(0,1))
qqnorm(D1C216.nlme, ~ ranef(.))
plot(D1C216.nlme,cex=0.7,adj=-1)
anova(D1C216.nlme,Terms=c(1:9))
#covariates jointly sig!

##############################################################################
##### Dataset 2 - day0 zero stays zero, others changed to 8 ##################
##############################################################################
#########Cex2#########

#No covariates
D2C21.nls <- nls(log2game ~ CEx2(day,C,R),data=Data2,start=c(C=2.2,R=0.93),trace=T)
#Solves to: 1.64, 0.93
D2C21.lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data2,start=c(C=1.8,R=0.95),na.action=na.pass)
#Solved for all!
D2C21.nlme<-nlme(D2C21.lis,random=C+R~1,verbose=T)
intervals(D2C21.nlme)
summary(D2C21.nlme)
D2C2.ran1 <- ranef(D2C21.nlme,augFrame=T)
plot(D2C2.ran1,form= C ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat followed by site and logpdens, maybe age
plot(D2C2.ran1,form= R ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#quite a few promising covs! Age, fail, pcat, site,pct?

#1st cov: age
D2C22.nlme <- update (D2C21.nlme,fixed=list(C ~ 1, R ~ age),start=c(1.675,0.93,0))
summary(D2C22.nlme)
D2C23.nlme <- update (D2C21.nlme,fixed=list(C ~ age, R ~ age),start=c(1.675,0,0.93,0))
summary(D2C23.nlme)
#age not sig on C
anova(D2C23.nlme,D2C22.nlme)
#better with age on R only
anova(D2C21.nlme,D2C22.nlme)
#age improves model
D2C2.ran2 <- ranef(D2C22.nlme,augFrame=T)
plot(D2C2.ran2,form= C ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat, site, and logpdens, maybe pct?
plot(D2C2.ran2,form= R.(Intercept) ~site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat, site, fail, pct, maybe mut2?

#2nd cov: site
D2C24.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site), R ~ age+factor(site)),start=c(1.675,0,0.93,0,
```

```
0))
summary(D2C24.nlme)
D2C25.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site), R ~ age),start=c(1.675,0,0.93,0))
summary(D2C25.nlme)
#Site not sig when used only on C
D2C26.nlme <- update (D2C21.nlme,fixed=list(C ~ 1, R ~ age+factor(site)),start=c(1.675,0.93,0,0))
summary(D2C26.nlme)
#Site sig on R
anova(D2C25.nlme,D2C24.nlme)
anova(D2C26.nlme,D2C24.nlme)
#better to have site on both parameters
anova(D2C24.nlme,D2C22.nlme)
#model better with site
anova(D2C24.nlme,Terms=c(2,5))
D2C2.ran3 <- ranef(D2C24.nlme,augFrame=T)
plot(D2C2.ran3,form= C.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat and logpdens?
plot(D2C2.ran3,form= R.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat, fail, logpdens

#3rd cov: logpdens
D2C27.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+logpdens, R ~ age+factor(site)+logpdens),start
=c(1.675,0,0,0.93,0,0,0))
summary(D2C27.nlme)
#neither is sig
D2C28.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+logpdens, R ~ age+factor(site)),start=c(1.675,
0,0,0.93,0,0))
summary(D2C28.nlme)
#now logpdens sig on C
D2C29.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site), R ~ age+factor(site)+logpdens),start=c(1.675,
0,0.93,0,0,0))
summary(D2C29.nlme)
anova(D2C28.nlme,D2C27.nlme)
anova(D2C29.nlme,D2C28.nlme)
anova(D2C29.nlme,D2C27.nlme)
#better to have logpdens on R
anova(D2C29.nlme,D2C24.nlme)
#better to include logpdens
D2C2.ran4 <- ranef(D2C29.nlme,augFrame=T)
plot(D2C2.ran4,form= C.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat and age?
plot(D2C2.ran4,form= R.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#pcat, fail?

#4th cov: age on C
D2C210.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+age, R ~ age+factor(site)+logpdens),start=c(1
.675,0,0,0.93,0,0,0))
summary(D2C210.nlme)
anova(D2C210.nlme,D2C29.nlme)
#better without age on C

#4th cov: logpdens on C
D2C211.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+logpdens, R ~ age+factor(site)+logpdens),star
t=c(1.675,0,0,0.93,0,0,0))
summary(D2C211.nlme)
anova(D2C211.nlme,D2C29.nlme)
```

```
#4th cov: fail on R
Data<-read.csv("Final_Data2_102.csv",header=T,sep=",")
attach(Data)
Data102_2<-transform(Data,log2game=log2(gamedens+1),pcat=as.factor(pcat),logpdens=log10(pdens0))
detach(Data)
Data102_2<-groupedData(log2game ~ day | subjectno, data=Data102_2)
D2C2102.lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data102_2,start=c(C=1.8,R=0.95),na.action=na.pass)
D2C212.nlme<-nlme(D2C2102.lis,random=C+R~1,verbose=T)
D2C213.nlme <- update (D2C212.nlme,fixed=list(C ~ factor(site), R ~ age+factor(site)+logpdens),start=c(1.67
5,0,0.93,0,0,0),na.action=na.omit)
D2C214.nlme <- update (D2C212.nlme,fixed=list(C ~ factor(site), R ~ age+factor(site)+logpdens+fail),start=c
(1.675,0,0.93,0,0,0,0),na.action=na.omit)
anova(D2C214.nlme,D2C215.nlme)

#4th cov: pcat
D2C215.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+factor(pcat), R ~ age+factor(site)+logpdens+f
actor(pcat)),start=c(1.675,0,0,0,0.93,0,0,0,0,0))
summary(D2C215.nlme)
#sig, except for diff btwn 3 and 1 on C as expected
D2C216.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site), R ~ age+factor(site)+logpdens+factor(pcat)),
start=c(1.675,0,0.93,0,0,0,0,0))
summary(D2C216.nlme)
#now pcat 2 not sig?
D2C217.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+factor(pcat), R ~ age+factor(site)+logpdens),
start=c(1.675,0,0,0,0.93,0,0,0))
summary(D2C217.nlme)
anova(D2C215.nlme,D2C216.nlme)
anova(D2C215.nlme,D2C217.nlme)
#better to have pcat on both parameters
anova(D2C215.nlme,D2C29.nlme)
#better with pcat
D2C2.ran5 <- ranef(D2C215.nlme,augFrame=T)
plot(D2C2.ran5,form= C.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#not much...fail?
plot(D2C2.ran5,form= R.(Intercept) ~ site+mut_3+pcat+fail+gender+logpdens+age+pct)
#not much...fail?

####test on reduced data.....
#5th cov: mut_3 on C
D2C218.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+factor(pcat)+factor(mut_3), R ~ age+factor(si
te)+logpdens+factor(pcat)),start=c(1.675,0,0,0,0,0.93,0,0,0,0,0),na.action=na.omit)
summary(D2C218.nlme)
D2C219.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+factor(pcat), R ~ age+factor(site)+logpdens+f
actor(pcat)+factor(mut_3)),start=c(1.675,0,0,0,0.93,0,0,0,0,0,0),na.action=na.omit)
summary(D2C219.nlme)
D2C220.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+factor(pcat)+factor(mut_3), R ~ age+factor(si
te)+logpdens+factor(pcat)+factor(mut_3)),start=c(1.675,0,0,0,0,0.93,0,0,0,0,0,0),na.action=na.omit)
summary(D2C220.nlme)
anova(D2C220.nlme,D2C218.nlme)
anova(D2C220.nlme,D2C219.nlme)
anova(D2C218.nlme,D2C219.nlme)
#best is with mutation on C only

D2C221.nlme <- update (D2C21.nlme,data=Data2R,fixed=list(C ~ factor(site)+factor(pcat), R ~ age+factor(site
)+logpdens+factor(pcat)),start=c(1.675,0,0,0,0.93,0,0,0,0,0))
anova(D2C221.nlme,D2C218.nlme)
#confirms that mutation doesnt improve model
```

```
####Trying few other things
#using pct as an ordered factor
Data2$pct2<-ordered(Data2$pct)
D2C2.ran5 <- ranef(D2C210.nlme,augFrame=T)
plot(D2C2.ran5,form= C.(Intercept) ~ pct2)
#nothing apparent
plot(D2C2.ran5,form= R.(Intercept) ~ pct2)
#could be 2 groupings?
Data2$pct2<-replace(Data2$pct2,Data2$pct2==2,1)
Data2$pct2<-replace(Data2$pct2,Data2$pct2==3,7)
D2C2.ran5 <- ranef(D2C210.nlme,augFrame=T)
plot(D2C2.ran5,form= C.(Intercept) ~ pct2)
plot(D2C2.ran5,form= R.(Intercept) ~ pct2)
#not much evidence (more for R)
D2C222.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site)+logpdens+factor(pcat), R ~ age+factor(site)+f
actor(pcat)+factor(pct2)),start=c(1.8777,0,0,0,0,0.9279,0,0,0,0,0))
summary(D2C222.nlme)
anova(D2C222.nlme,D2C210.nlme)
#does nothing

#checking interaction with mut and site on C
D2C21R.lis <- nlsList(log2game ~ CEx2(day,C,R),data=Data2R,start=c(C=2,R=0.95),na.action=na.pass)
#Solved for all!
D2C21R.nlme<-nlme(D2C21R.lis,random=C+R~1,verbose=T)
D2C22R.nlme <- update (D2C21R.nlme,fixed=list(C ~ factor(site)+factor(pcat), R ~ age+factor(site)+factor(pc
at)+logpdens),start=c(1.8777,0,0,0,0.9279,0,0,0,0,0))
D2C23R.nlme <- update (D2C21R.nlme,fixed=list(C ~ factor(site)*factor(mut_3)+factor(pcat), R ~ age+factor(s
ite)+factor(pcat)+logpdens),start=c(1.8777,0,0,0,0,0,0.9279,0,0,0,0,0))
summary(D2C23R.nlme)
D2C24R.nlme <- update (D2C21R.nlme,fixed=list(C ~ factor(site)+factor(pcat), R ~ age+factor(site)*factor(mu
t_3)+factor(pcat)+logpdens),start=c(1.8777,0,0,0,0.9279,0,0,0,0,0,0,0))
D2C25R.nlme <- update (D2C21R.nlme,fixed=list(C ~ factor(site)*factor(mut_3)+factor(pcat), R ~ age+factor(s
ite)*factor(mut_3)+factor(pcat)+logpdens),start=c(1.8777,0,0,0,0,0,0.9279,0,0,0,0,0,0,0))
anova(D2C25R.nlme,D2C23R.nlme)
#better on one
anova(D2C23R.nlme,D2C24R.nlme)
#better on C
anova(D2C23R.nlme,D2C22R.nlme)
#no improvement!

#Final model - diagnostics
plot(D2C215.nlme, log2game ~ fitted(.), abline=c(0,1))
qqnorm(D2C215.nlme, ~ ranef(.))
plot(D2C215.nlme,cex=0.7,adj=-1)
```

D.2.3 Model plots

```
##########################################
Fit same versions to different data
##########################################
D1Final1.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens+factor(pcat), R ~ age+factor(s
ite)+factor(pcat)),start=c(1.8777,0,0,0,0,0,0.9279,0,0,0,0))
summary(D1Final1.nlme)
D2Final1.nlme <- update (D2C21.nlme,fixed=list(C ~ age+factor(site)+logpdens+factor(pcat), R ~ age+factor(s
ite)+factor(pcat)),start=c(1.8777,0,0,0,0,0,0.9279,0,0,0,0))
```

```
summary(D2Final1.nlme)

D1Final2.nlme <- update (D1C21.nlme,fixed=list(C ~ factor(site)+factor(pcat), R ~ factor(site)+factor(pcat)
+age+logpdens),start=c(1.8777,0,0,0,0.9279,0,0,0,0,0))
summary(D1Final2.nlme)
D2Final2.nlme <- update (D2C21.nlme,data=Data2,fixed=list(C ~ factor(site)+factor(pcat), R ~ factor(site)+f
actor(pcat)+age+logpdens),start=c(1.8777,0,0,0,0.9279,0,0,0,0,0))
summary(D2Final2.nlme)

#without pcat
D1M1_nopcat.nlme <- update (D1C21.nlme,fixed=list(C ~ age+factor(site)+logpdens, R ~ age+factor(site)),star
t=c(1.8777,0,0,0,0.9279,0,0))
summary(D1M1_nopcat.nlme)
D2M1_nopcat.nlme <- update (D2C21.nlme,fixed=list(C ~ age+factor(site)+logpdens, R ~ age+factor(site)),star
t=c(1.8777,0,0,0,0.9279,0,0))
summary(D2M1_nopcat.nlme)

D1M2_nopcat.nlme <- update (D1C21.nlme,fixed=list(C ~ factor(site), R ~ age+factor(site)+logpdens),start=c(
1.675,0,0.93,0,0,0))
summary(D1M2_nopcat.nlme)
D2M2_nopcat.nlme <- update (D2C21.nlme,fixed=list(C ~ factor(site), R ~ age+factor(site)+logpdens),start=c(
1.675,0,0.93,0,0,0))
summary(D2M2_nopcat.nlme)

##########################
Fit typical pcat curves
##########################

####by subject####
attach(Data2)
x1<-day[subjectno=="M02_151"]
y1<-gamedens[subjectno=="M02_151"]
x2<-day[subjectno=="M02_097"]
y2<-gamedens[subjectno=="M02_097"]
x3<-day[subjectno=="M02_094"]
y3<-gamedens[subjectno=="M02_094"]

par(mfrow=c(3,1))
plot(x1,y1,xlab="Day",ylab="Log2 Gametocytes",main="Typical curves for patient category 1",ylim=c(0,900),xl
im=c(0,42),col="green", pch=19,type="b")
plot(x3,y3,xlab="Day",ylab="Log2 Gametocytes",main="Typical curves for patient category 2",ylim=c(0,900),xl
im=c(0,42), col="red", pch=23,type="b")
plot (x2,y2,xlab="Day",ylab="Log2 Gametocytes",main="Typical curves for patient category 3",ylim=c(0,900),x
lim=c(0,42), col="blue", pch=22,type="b")

##########################
Fit population curves
##########################
x <- seq(0,70, length=100000)
Av_age_Moz <- 17
Av_age_Mpm <- 18
###low and high refer to the 25th and 75th percentiles
Low_age_Moz <- 5
High_age_Moz <- 29
Low_age_Mpm <- 11
High_age_Mpm <- 28.5
Av_pdens <- 4.5
```

```
Low_pdens <- 4.2
High_pdens <- 4.8

#Data2, Model 2
CMoz1_4 <- 2.1077127
CMoz2_4 <- 2.1077127-0.0590195
CMoz3_4 <- 2.1077127-0.491024
RMoz1_4 <- 0.8845865+(0.0001775*Av_age_Moz)+(0.006104*Av_pdens)
RMoz2_4 <- 0.8845865+0.0105655+(0.0001775*Av_age_Moz)+(0.006104*Av_pdens)
RMoz3_4 <- 0.8845865+0.0133872+(0.0001775*Av_age_Moz)+(0.006104*Av_pdens)

CMpm1_4 <- 2.1077127-0.4987982
CMpm2_4 <- 2.1077127-0.4987982-0.0590195
CMpm3_4 <- 2.1077127-0.4987982-0.491024
RMpm1_4 <- 0.8845865+0.0153124+(0.0001775*Av_age_Mpm)+(0.006104*Av_pdens)
RMpm2_4 <- 0.8845865+0.0153124+0.0105655+(0.0001775*Av_age_Mpm)+(0.006104*Av_pdens)
RMpm3_4 <- 0.8845865+0.0153124+0.0133872+(0.0001775*Av_age_Mpm)+(0.006104*Av_pdens)

y22Moz1 <- (CMoz1_4*x)*(RMoz1_4^x)
y22Moz2 <- (CMoz2_4*x)*(RMoz2_4^x)
y22Moz3 <- (CMoz3_4*x)*(RMoz3_4^x)
y22Mpm1 <- (CMpm1_4*x)*(RMpm1_4^x)
y22Mpm2 <- (CMpm2_4*x)*(RMpm2_4^x)
y22Mpm3 <- (CMpm3_4*x)*(RMpm3_4^x)

#Popn curves by site and pcat
par(mfcol=c(1,2))
plot(c(0,70),c(0,10),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Moz1,type="l", lwd=2, col=2)
lines(x,y22Moz2, type="l", lwd=2, col=3)
lines(x,y22Moz3, type="l", lwd=2, col=4)
legend(10,2,cex=0.75, legend=c('Cat1', 'Cat2', 'Cat3'),
    lty=1,col=c(2,3,4))
title('Mozambique')

plot(c(0,70),c(0,10),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Mpm1,type="l", lwd=2, col=2)
lines(x,y22Mpm2, type="l", lwd=2, col=3)
lines(x,y22Mpm3, type="l", lwd=2, col=4)
legend(10,2,cex=0.75, legend=c('Cat1', 'Cat2', 'Cat3'),
   lty=1, col=c(2,3,4))
title('Mpumulanga')

#################################################################
#Data2, Model 2 without pcat
# showing age effect (young vs old) within site for avg pdens
CMoz1_5 <- 2.0593839
CMoz2_5 <- 2.0593839
RMoz1_5 <- 0.8867244+(0.000153*Low_age_Moz)+(0.0063258*Av_pdens)
RMoz2_5 <- 0.8867244+(0.000153*High_age_Moz)+(0.0063258*Av_pdens)

CMpm1_5 <- 2.0593839-0.5405995
CMpm2_5 <- 2.0593839-0.5405995
RMpm1_5 <- 0.8867244+0.0164081+(0.000153*Low_age_Mpm)+(0.0063258*Av_pdens)
RMpm2_5 <- 0.8867244+0.0164081+(0.000153*High_age_Mpm)+(0.0063258*Av_pdens)

y22MozY <- (CMoz1_5*x)*(RMoz1_5^x)
```

```
y22MozO <- (CMoz2_5*x)*(RMoz2_5^x)
y22MpmY <- (CMpm1_5*x)*(RMpm1_5^x)
y22MpmO <- (CMpm2_5*x)*(RMpm2_5^x)

#Popn curves by site
par(mfcol=c(1,2))
plot(c(0,70),c(0,10),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22MozY,type="l", lwd=2, lty=1,col=2)
lines(x,y22MozO, type="l", lwd=2, lty=3,col=2)
legend(10,2,cex=0.75, legend=c('Young', 'Old'),
    lty=c(1,3),col=2)
text(x=50, y=10,'Mozambique')
title('Age effect for average logpdens')

plot(c(0,70),c(0,10),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22MpmY,type="l", lwd=2, lty=1,col=4)
lines(x,y22MpmO, type="l", lwd=2, lty=3,col=4)
legend(10,2,cex=0.75, legend=c('Young', 'Old'),
    lty=c(1,3), col=4)
text(x=50, y=10,'Mpumulanga')
title('Age effect for average logpdens')

#Showing site effect for median age and avg logpdens
CMoz_5 <- 2.0593839
CMpm_5 <- 2.0593839-0.5405995
RMoz_5 <- 0.8867244+(0.000153*Av_age_Moz)+(0.0063258*Av_pdens)
RMpm_5 <- 0.8867244+0.0164081+(0.000153*Av_age_Mpm)+(0.0063258*Av_pdens)

y22Moz <- (CMoz_5*x)*(RMoz_5^x)
y22Mpm <- (CMpm_5*x)*(RMpm_5^x)

plot(c(0,70),c(0,10),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Moz,type="l", lwd=2, lty=1,col=2)
lines(x,y22Mpm, type="l", lwd=2, lty=3,col=4)
legend(10,2,cex=0.75, legend=c('Mozambique', 'Mpumulanga'),
    lty=c(1,3), col=c(2,4))
title('Site effect for median age & avg logpdens')

#Showing pdens effect for median age, within site
CMoz1_6 <- 2.0593839
CMoz2_6 <- 2.0593839
CMpm1_6 <- 2.0593839-0.5405995
CMpm2_6 <- 2.0593839-0.5405995
RMoz1_6 <- 0.8867244+(0.000153*Av_age_Moz)+(0.0063258*Low_pdens)
RMoz2_6 <- 0.8867244+(0.000153*Av_age_Moz)+(0.0063258*High_pdens)
RMpm1_6 <- 0.8867244+0.0164081+(0.000153*Av_age_Mpm)+(0.0063258*Low_pdens)
RMpm2_6 <- 0.8867244+0.0164081+(0.000153*Av_age_Mpm)+(0.0063258*High_pdens)

y22MozLP <- (CMoz1_6*x)*(RMoz1_6^x)
y22MozHP <- (CMoz2_6*x)*(RMoz2_6^x)
y22MpmLP <- (CMpm1_6*x)*(RMpm1_6^x)
y22MpmHP <- (CMpm2_6*x)*(RMpm2_6^x)

#by site
par(mfcol=c(1,2))
plot(c(0,70),c(0,10),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22MozLP,type="l", lwd=2, lty=1,col=2)
```

```
lines(x,y22MozHP,type="l", lwd=2, lty=3,col=2)
legend(10,2,cex=0.75, legend=c('Low', 'High'),
    lty=c(1,3),col=2)
text(x=50, y=10,'Mozambique')
title('Logged pdens effect with median age')

plot(c(0,70),c(0,10),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22MpmLP, type="l", lwd=2, lty=1,col=4)
lines(x,y22MpmHP, type="l", lwd=2, lty=3,col=4)
legend(10,2,cex=0.75, legend=c('Low', 'High'),
    lty=c(1,3), col=c(4))
text(x=50, y=10,'Mpumulanga')
title('Logged pdens effect with median age')

#########################################################################
#####versus observed points###########
#Popn Curves vs observed (mutations) by pcat
par(mfcol=c(2,3))
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Moz1,type="l", lwd=2, lty=1,col=2)
points(Data2$day[Data2$pcat==1 & Data2$site=="Moz" & Data2$mut_3=="Sensitive"], Data2$log2game[Data2$pcat==
1 & Data2$site=="Moz" & Data2$mut_3=="Sensitive"], pch=22, col="black")
points(Data2$day[Data2$pcat==1 & Data2$site=="Moz" & Data2$mut_3=="Resistant"], Data2$log2game[Data2$pcat==
1 & Data2$site=="Moz" & Data2$mut_3=="Resistant"], pch=20, col="green")
legend(50,10,cex=0.75, legend=c('Sensitive', 'Resistant'),
   pch=c(22,20), col=c("black","green"))
title('Mozambique Cat 1: fitted vs observed (mutation)')

plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Mpm1,type="l", lwd=2, lty=1,col=4)
points(Data2$day[Data2$pcat==1 & Data2$site=="Mpumulanga" & Data2$mut_3=="Sensitive"], Data2$log2game[Data2
$pcat==1 & Data2$site=="Mpumulanga" & Data2$mut_3=="Sensitive"], pch=22, col="black")
points(Data2$day[Data2$pcat==1 & Data2$site=="Mpumulanga" & Data2$mut_3=="Resistant"], Data2$log2game[Data2
$pcat==1 & Data2$site=="Mpumulanga" & Data2$mut_3=="Resistant"], pch=20, col="green")
legend(50,10,cex=0.75, legend=c('Sensitive', 'Resistant'),
   pch=c(22,20), col=c("black","green"))
title('Mpumulanga Cat 1: fitted vs observed (mutation)')

plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Moz2,type="l", lwd=2, lty=1,col=2)
points(Data2$day[Data2$pcat==2 & Data2$site=="Moz" & Data2$mut_3=="Sensitive"], Data2$log2game[Data2$pcat==
2 & Data2$site=="Moz" & Data2$mut_3=="Sensitive"], pch=22, col="black")
points(Data2$day[Data2$pcat==2 & Data2$site=="Moz" & Data2$mut_3=="Resistant"], Data2$log2game[Data2$pcat==
2 & Data2$site=="Moz" & Data2$mut_3=="Resistant"], pch=20, col="green")
legend(50,10,cex=0.75, legend=c('Sensitive', 'Resistant'),
   pch=c(22,20), col=c("black","green"))
title('Mozambique Cat 2: fitted vs observed (mutation)')

plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Mpm2,type="l", lwd=2, lty=1,col=4)
points(Data2$day[Data2$pcat==2 & Data2$site=="Mpumulanga" & Data2$mut_3=="Sensitive"], Data2$log2game[Data2
$pcat==2 & Data2$site=="Mpumulanga" & Data2$mut_3=="Sensitive"], pch=22, col="black")
points(Data2$day[Data2$pcat==2 & Data2$site=="Mpumulanga" & Data2$mut_3=="Resistant"], Data2$log2game[Data2
$pcat==2 & Data2$site=="Mpumulanga" & Data2$mut_3=="Resistant"], pch=20, col="green")
legend(50,10,cex=0.75, legend=c('Sensitive', 'Resistant'),
   pch=c(22,20), col=c("black","green"))
title('Mpumulanga Cat 2: fitted vs observed (mutation)')
```

```
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Moz3,type="l", lwd=2, lty=1,col=2)
points(Data2$day[Data2$pcat==3 & Data2$site=="Moz" & Data2$mut_3=="Sensitive"], Data2$log2game[Data2$pcat==
3 & Data2$site=="Moz" & Data2$mut_3=="Sensitive"], pch=22, col="black")
points(Data2$day[Data2$pcat==3 & Data2$site=="Moz" & Data2$mut_3=="Resistant"], Data2$log2game[Data2$pcat==
3 & Data2$site=="Moz" & Data2$mut_3=="Resistant"], pch=20, col="green")
legend(50,10,cex=0.75, legend=c('Sensitive', 'Resistant'),
   pch=c(22,20), col=c("black","green"))
title('Mozambique Cat 3: fitted vs observed (mutation)')

plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
lines(x,y22Mpm3,type="l", lwd=2, lty=1,col=4)
points(Data2$day[Data2$pcat==3 & Data2$site=="Mpumulanga" & Data2$mut_3=="Sensitive"], Data2$log2game[Data2
$pcat==3 & Data2$site=="Mpumulanga" & Data2$mut_3=="Sensitive"], pch=22, col="black")
points(Data2$day[Data2$pcat==3 & Data2$site=="Mpumulanga" & Data2$mut_3=="Resistant"], Data2$log2game[Data2
$pcat==3 & Data2$site=="Mpumulanga" & Data2$mut_3=="Resistant"], pch=20, col="green")
legend(50,10,cex=0.75, legend=c('Sensitive', 'Resistant'),
   pch=c(22,20), col=c("black","green"))
title('Mpumulanga Cat 3: fitted vs observed (mutation)')

##############################################################################################
################## Plotting pop curve vs fitted lines by subject
##############################################################################################
coef <- coef(D2Final2.nlme,augFrame=T,data=Data2, which=c(1,5:8,14,22),omitGroupingFactor=TRUE)
Mozcat1<- coef[coef$pcat=="1" & coef$site=="Moz",]
Mpmcat1<- coef[coef$pcat=="1" & coef$site=="Mpumulanga",]
Mozcat2<- coef[coef$pcat=="2" & coef$site=="Moz",]
Mpmcat2<- coef[coef$pcat=="2" & coef$site=="Mpumulanga",]
Mozcat3<- coef[coef$pcat=="3" & coef$site=="Moz",]
Mpmcat3<- coef[coef$pcat=="3" & coef$site=="Mpumulanga",]

par(mfrow=c(2,3))
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:24){
C[i]<-Mozcat1[i,1]
R[i]<-Mozcat1[i,5]+(Mozcat1[i,9]*Mozcat1$age)+(Mozcat1[i,10]*Mozcat1$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1,lty=3) }
lines(x,y22Moz1,col="blue",lwd=2)
legend(35,12,cex=0.85, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mozambique, Cat 1')
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:3){
C[i]<-Mozcat2[i,1]+Mozcat2[i,3]
R[i]<-Mozcat2[i,5]+Mozcat2[i,7]+(Mozcat2[i,9]*Mozcat2$age)+(Mozcat2[i,10]*Mozcat2$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1) }
lines(x,y22Moz2,col="blue",lwd=2)
legend(35,12,cex=0.85, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mozambique, Cat 2')
```

```
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:4){
C[i]<-Mozcat3[i,1]+Mozcat3[i,4]
R[i]<-Mozcat3[i,5]+Mozcat3[i,8]+(Mozcat3[i,9]*Mozcat3$age)+(Mozcat3[i,10]*Mozcat3$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1) }
lines(x,y22Moz3,col="blue",lwd=2)
legend(35,12,cex=0.85, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mozambique, Cat 3')
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:50){
C[i]<-Mpmcat1[i,1]+Mpmcat1[i,2]
R[i]<-Mpmcat1[i,5]+Mpmcat1[i,6]+(Mpmcat1[i,9]*Mpmcat1$age)+(Mpmcat1[i,10]*Mpmcat1$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1,lty=2) }
lines(x,y22Mpm1,col="blue",lwd=2)
legend(35,12,cex=0.85, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mpumulanga, Cat 1')
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:12){
C[i]<-Mpmcat2[i,1]+Mpmcat2[i,2]+Mpmcat2[i,3]
R[i]<-Mpmcat2[i,5]+Mpmcat2[i,6]+Mpmcat2[i,7]+(Mpmcat2[i,9]*Mpmcat2$age)+(Mpmcat2[i,10]*Mpmcat2$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1) }
lines(x,y22Mpm2,col="blue",lwd=2)
legend(35,12,cex=0.85, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mpumulanga, Cat 2')
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:12){
C[i]<-Mpmcat3[i,1]+Mpmcat3[i,2]+Mpmcat3[i,4]
R[i]<-Mpmcat3[i,5]+Mpmcat3[i,6]+Mpmcat3[i,8]+(Mpmcat3[i,9]*Mpmcat3$age)+(Mpmcat3[i,10]*Mpmcat3$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1) }
lines(x,y22Mpm3,col="blue",lwd=2)
legend(35,12,cex=0.85, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mpumulanga, Cat 3')

#Cat 1
par(mfrow=c(2,1))
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:24){
C[i]<-Mozcat1[i,1]
```

```
R[i]<-Mozcat1[i,5]+(Mozcat1[i,9]*Mozcat1$age)+(Mozcat1[i,10]*Mozcat1$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1,lty=3) }
lines(x,y22Moz1,col="blue",lwd=2)
legend(50,10,cex=0.75, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mozambique, Cat 1')
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:50){
C[i]<-Mpmcat1[i,1]+Mpmcat1[i,2]
R[i]<-Mpmcat1[i,5]+Mpmcat1[i,6]+(Mpmcat1[i,9]*Mpmcat1$age)+(Mpmcat1[i,10]*Mpmcat1$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1,lty=2) }
lines(x,y22Mpm1,col="blue",lwd=2)
legend(50,10,cex=0.75, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mpumulanga, Cat 1')

#Cat 2
par(mfrow=c(2,1))
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:3){
C[i]<-Mozcat2[i,1]+Mozcat2[i,3]
R[i]<-Mozcat2[i,5]+Mozcat2[i,7]+(Mozcat2[i,9]*Mozcat2$age)+(Mozcat2[i,10]*Mozcat2$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1) }
lines(x,y22Moz2,col="blue",lwd=2)
legend(50,10,cex=0.75, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mozambique, Cat 2')
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:12){
C[i]<-Mpmcat2[i,1]+Mpmcat2[i,2]+Mpmcat2[i,3]
R[i]<-Mpmcat2[i,5]+Mpmcat2[i,6]+Mpmcat2[i,7]+(Mpmcat2[i,9]*Mpmcat2$age)+(Mpmcat2[i,10]*Mpmcat2$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1) }
lines(x,y22Mpm2,col="blue",lwd=2)
legend(50,13,cex=0.75, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mpumulanga, Cat 2')

#Cat 3
par(mfrow=c(2,1))
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:4){
C[i]<-Mozcat3[i,1]+Mozcat3[i,4]
R[i]<-Mozcat3[i,5]+Mozcat3[i,8]+(Mozcat3[i,9]*Mozcat3$age)+(Mozcat3[i,10]*Mozcat3$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1) }
```

```
lines(x,y22Moz3,col="blue",lwd=2)
legend(50,10,cex=0.75, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mozambique, Cat 3')
C<-NULL
R<-NULL
Y<-NULL
plot(c(0,70),c(0,14),type='n',xlab='Days', ylab='Log2Gamedens')
for (i in 1:12){
C[i]<-Mpmcat3[i,1]+Mpmcat3[i,2]+Mpmcat3[i,4]
R[i]<-Mpmcat3[i,5]+Mpmcat3[i,6]+Mpmcat3[i,8]+(Mpmcat3[i,9]*Mpmcat3$age)+(Mpmcat3[i,10]*Mpmcat3$logpdens)
Y<-(C[i]*x)*(R[i]^x)
lines(x,Y,col="red",lwd=1) }
lines(x,y22Mpm3,col="blue",lwd=2)
legend(50,10,cex=0.75, legend=c('Subject', 'Population'),lty=1, col=c("red","blue"))
title('Mpumulanga, Cat 3')
```

D.2.4 Final models

Final Model for Data1

```
Nonlinear mixed-effects model fit by maximum likelihood
  Model: log2game ~ CEx2(day, C, R)
 Data: Data1
       AIC      BIC    logLik
  1821.079 1883.965 -895.5397

Random effects:
 Formula: list(C ~ 1, R ~ 1)
 Level: subjectno
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
C.(Intercept) 0.53074361 C.(In)
R.(Intercept) 0.01047509 -0.74
Residual      1.11997104

Fixed effects: list(C ~ age + factor(site) + logpdens + factor(pcat), R ~ age + factor(site) + factor(pcat))
                              Value Std.Error  DF   t-value p-value
C.(Intercept)              1.3113815 0.3825433 376   3.42806  0.0007
C.age                     -0.0068819 0.0038745 376  -1.77621  0.0765
C.factor(site)Mpumulanga  -0.4149698 0.1409572 376  -2.94394  0.0034
C.logpdens                 0.2468307 0.0805601 376   3.06393  0.0023
C.factor(pcat)2           -0.1999889 0.1817571 376  -1.10031  0.2719
C.factor(pcat)3           -0.6015263 0.1750753 376  -3.43581  0.0007
R.(Intercept)              0.9058749 0.0033955 376 266.78745  0.0000
R.age                      0.0003466 0.0000897 376   3.86250  0.0001
R.factor(site)Mpumulanga   0.0133747 0.0033120 376   4.03828  0.0001
R.factor(pcat)2            0.0131051 0.0040039 376   3.27309  0.0012
R.factor(pcat)3            0.0180561 0.0040682 376   4.43832  0.0000
 Correlation:
                          C.(In) C.age  C.f()M C.lgpd C.f()2 C.f()3 R.(In)
C.age                     -0.162
C.factor(site)Mpumulanga  -0.230 -0.074
C.logpdens                -0.923 -0.054 -0.004
C.factor(pcat)2           -0.100  0.163 -0.041 -0.013
C.factor(pcat)3            0.046 -0.114 -0.100 -0.089  0.171
R.(Intercept)             -0.290  0.427  0.475 -0.016  0.231  0.078
R.age                      0.187 -0.771  0.072 -0.026 -0.121  0.088 -0.523
R.factor(site)Mpumulanga   0.180  0.069 -0.783  0.000  0.035  0.086 -0.596
```

```
R.factor(pcat)2           0.085 -0.128  0.037  0.008 -0.778 -0.148 -0.302
R.factor(pcat)3          -0.002  0.084  0.087  0.034 -0.140 -0.766 -0.099
                         R.age  R.f()M R.f()2
C.age
C.factor(site)Mpumulanga
C.logpdens
C.factor(pcat)2
C.factor(pcat)3
R.(Intercept)
R.age
R.factor(site)Mpumulanga -0.118
R.factor(pcat)2           0.149 -0.046
R.factor(pcat)3          -0.110 -0.111  0.190

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.7757339 -0.3932691  0.0000000  0.4465723  3.8410642

Number of Observations: 489
Number of Groups: 103
```

Final Model for Data2

```
Nonlinear mixed-effects model fit by maximum likelihood
  Model: log2game ~ CEx2(day, C, R)
 Data: Data2
       AIC      BIC    logLik
  2727.078 2790.793 -1349.539

Random effects:
 Formula: list(C ~ 1, R ~ 1)
 Level: subjectno
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
C.(Intercept) 0.48540306 C.(In)
R.(Intercept) 0.01005877 -0.745
Residual      1.37034391

Fixed effects: list(C ~ factor(site) + factor(pcat), R ~ factor(site) + factor(pcat) + age + logpdens)
                             Value  Std.Error  DF   t-value p-value
C.(Intercept)             2.1077127 0.11413864 588  18.46625  0.0000
C.factor(site)Mpumulanga -0.4987982 0.12901755 588  -3.86613  0.0001
C.factor(pcat)2          -0.0590195 0.16513862 588  -0.35739  0.7209
C.factor(pcat)3          -0.4910240 0.15786186 588  -3.11047  0.0020
R.(Intercept)             0.8845865 0.00827024 588 106.96024  0.0000
R.factor(site)Mpumulanga  0.0153124 0.00306867 588   4.98993  0.0000
R.factor(pcat)2           0.0105655 0.00378460 588   2.79171  0.0054
R.factor(pcat)3           0.0133872 0.00390716 588   3.42632  0.0007
R.age                     0.0001775 0.00005416 588   3.27740  0.0011
R.logpdens                0.0061040 0.00175054 588   3.48691  0.0005
 Correlation:
                         C.(In) C.f()M C.f()2 C.f()3 R.(In) R.f()M R.f()2
C.factor(site)Mpumulanga -0.801
C.factor(pcat)2          -0.231 -0.022
C.factor(pcat)3          -0.184 -0.095  0.181
R.(Intercept)            -0.268  0.200  0.057  0.061
R.factor(site)Mpumulanga  0.624 -0.777  0.016  0.071 -0.243
R.factor(pcat)2           0.188  0.017 -0.772 -0.146 -0.077 -0.027
```

```
R.factor(pcat)3            0.139  0.067 -0.136 -0.755 -0.018 -0.079  0.175
R.age                      0.002  0.010  0.001  0.002  0.004 -0.070  0.095
R.logpdens                 0.012  0.004  0.002 -0.014 -0.936 -0.009 -0.016
                          R.f()3 R.age
C.factor(site)Mpumulanga
C.factor(pcat)2
C.factor(pcat)3
R.(Intercept)
R.factor(site)Mpumulanga
R.factor(pcat)2
R.factor(pcat)3
R.age                     -0.076
R.logpdens                -0.033 -0.142

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.8382780 -0.4747489  0.0000000  0.4852039  3.7986408

Number of Observations: 700
Number of Groups: 103
```

Model refitted without patient category

```
Nonlinear mixed-effects model fit by maximum likelihood
  Model: log2game ~ CEx2(day, C, R)
 Data: Data2
       AIC      BIC    logLik
  2742.319 2787.829 -1361.159

Random effects:
 Formula: list(C ~ 1, R ~ 1)
 Level: subjectno
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
C.(Intercept) 0.52249608 C.(In)
R.(Intercept) 0.01172114 -0.751
Residual      1.36493530

Fixed effects: list(C ~ factor(site), R ~ age + factor(site) + logpdens)
                            Value  Std.Error  DF  t-value p-value
C.(Intercept)             2.0593839 0.11496037 592 17.91386  0.0000
C.factor(site)Mpumulanga -0.5405995 0.13471200 592 -4.01300  0.0001
R.(Intercept)             0.8867244 0.00892207 592 99.38547  0.0000
R.age                     0.0001530 0.00005759 592  2.65610  0.0081
R.factor(site)Mpumulanga  0.0164081 0.00332349 592  4.93699  0.0000
R.logpdens                0.0063258 0.00189480 592  3.33850  0.0009
 Correlation:
                          C.(In) C.f()M R.(In) R.age  R.f()M
C.factor(site)Mpumulanga -0.853
R.(Intercept)            -0.258  0.210
R.age                    -0.001  0.008  0.013
R.factor(site)Mpumulanga  0.665 -0.779 -0.251 -0.072
R.logpdens                0.010  0.000 -0.941 -0.144 -0.009
Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.8393605 -0.4771249  0.0000000  0.4777625  3.7520298
Number of Observations: 700
Number of Groups: 103
```