Modeling Victoria's Injection Drug Users

by

Ryan Alexander Stone
B.Sc., University of Victoria, 2007
A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in the Department of Mathematics and Statistics

Supervisory Committee

Dr. Laura Cowen, Supervisor
(Department of Mathematics and Statistics)

Dr. Farouk Nathoo, Departmental Member
(Department of Mathematics and Statistics)

Dr. Eric Roth, Outside Member
(Department of Anthropology)

Supervisory Committee
Dr. Laura Cowen, Supervisor
(Department of Mathematics and Statistics)

Dr. Farouk Nathoo, Departmental Member
(Department of Mathematics and Statistics)

Dr. Eric Roth, Outside Member
(Department of Anthropology)

ABSTRACT

The objective of this thesis is to examine random effect models applied to binary data. I will use classical and Bayesian inference to fit generalized linear mixed models to a specific data set. The data analyzed in this thesis comes from a study examining the injection practices of needle exchange clientele in Victoria, B.C. focusing on their risk networks. First, I will examine the application of social network analysis to the study of injection drug use, focusing on issues of gender, norms, and the problem of hidden populations. Next the focus will be on random effect models, where I will provide some background and a few examples pertaining to generalized linear mixed models (GLMMs). After GLMMs, I will discuss the nature of the injection drug use study and the data which will then be analyzed using a GLMM. Lastly, I will provide a discussion about my results of the GLMM analysis along with a summary of the injection practices of the needle exchange clientele.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank everyone who was involved in the completion of this thesis. There were several people along the course of five years who I owe a lot of thanks. Starting with the most important person who helped me with the completion of this thesis, Laura Cowen. I would like to thank Laura for her patience and motivation to get me graduated. I want to especially thank her for taking me on as a student and having an extreme set of patience to grind this out with me. Without her help there is no way I would have done this. Eric Roth for all the work he has done with me, believing in me, and giving me the different opportunities to get me to where I am. Farouk Nathoo for taking the time to assist me when Laura was away. Coming into this in the middle must have been tough. I think it is extremely important to have good supervisors and I was blessed with three amazing ones. Each with their own set of strengths and values but all extremely important in helping me along the way. I would also like to thank every fellow student who helped me in every class and with my work on this thesis. I would like to thank Rabih Saab for putting up with me in the office. For always being there for me, listening to all my stories and hearing about my crazy life. You are a great friend Rabih and wish you the best. To all my friends and family, you have created the man I am today, I am thankful for all your guidance and support.
Ryan Stone

Chapter 1

Introduction

The objective of this thesis is to examine random effect models applied to binary data. These models will be analyzed with classical and Bayesian inference using generalized linear mixed models to a specific data set. The data analyzed in this thesis come from a study examining the injection practices of needle exchange clientele in Victoria, B.C. focusing on their risk networks. First, I will examine the application of social network analysis to the study of injection drug use, focusing on issues of gender, norms, and the problem of hidden populations. Next the focus will be on random effect models, where I will provide some background and a few examples pertaining to generalized linear mixed models (GLMMs). I will then introduce the nature of the injection drug use study and the data which will be analyzed using a GLMM. Lastly, I will discuss my results of the GLMM analysis along with a summary of the injection practices of the needle exchange clientele.

1.1 Network History

One of the first known types of networks was created by Euler in 1736 for the Konigsberg Bridge Problem (Newman et al., 2006). This problem was a mathematical riddle about a way to cross seven bridges that connected two land masses. The problem was, is there a way to cross the city using all seven bridges without crossing a bridge more than once? Euler showed that this particular problem had no solution by using a graph. A graph is defined by a set of vertices or nodes that are possibly connected by edges. (Gross and Yellen, 2004)

The characteristics of social ties, relationships such as companionship and inter-

action (Laireiter and Baumann, 1992), were reported in several published studies stemming from the 1920's. However, an important contribution to network analysis was made in 1934 by Jacob L. Moreno. Sociograms were Moreno's way of depicting a relationship on paper. A sociogram is made by drawing nodes, representing people connected by lines forming the relationship (Luke and Harris, 2007). At around the same time that Moreno was developing sociograms, Harvard researchers were being influenced by a social anthropologist named Alfred Radcliffe-Brown. They came up with the idea that networks were comprised of several smaller closely related sub groups, called cliques (Scott, 2000). Another small group of researchers from the Department of Social Anthropology at Manchester University was also being influenced by Radcliffe-Brown (Scott, 2000). They in turn looked at his ideas of cohesion and integration among a social structure and decided to focus on conflict and change (Scott, 2000). These three ideas came together in the 1960's and the early 1970's to form contemporary social network analysis (Scott, 2000).

Instead of centering the analysis on an individual, a different view or "level" of analysis can be focused on the individual's network. A network is comprised of links or bonds between people that form a group (Friedman and Aral, 2001). A social network can be looked at as a web of social links or bonds that compose a group of people. These links or bonds are any type of social tie that can happen between two people (Friedman et al., 1999). Some examples of these ties are: friendship, family, co-workers, or classmates. Social networks play an important role in determining the risk behaviour and characteristics of People Who Use Injection Drugs (PWUID). According to De et al. (2007) characteristics of social networks have been linked to the initiation, continuation, and cessation of drug consumption. These networks are formed by the links that relationships provide amongst people. A link provides a direct route through which information, education, money, drugs, and infections can be passed.

Another link or bond that forms a network is the sharing of injection drugs; this would form a risk network. A risk network is one that can potentially spread infection between two people (Friedman and Aral, 2001). Sexual risk networks occur when people have unprotected sex. Each risk is not unique to a single network, there can be multiple risks associated with a single network. For example, sex partners that also share drugs results in a network with more than one risk.

Social networks can be viewed in two different ways. First, if we look at the mindset of a single person, this is known as egocentric. To view a network by only

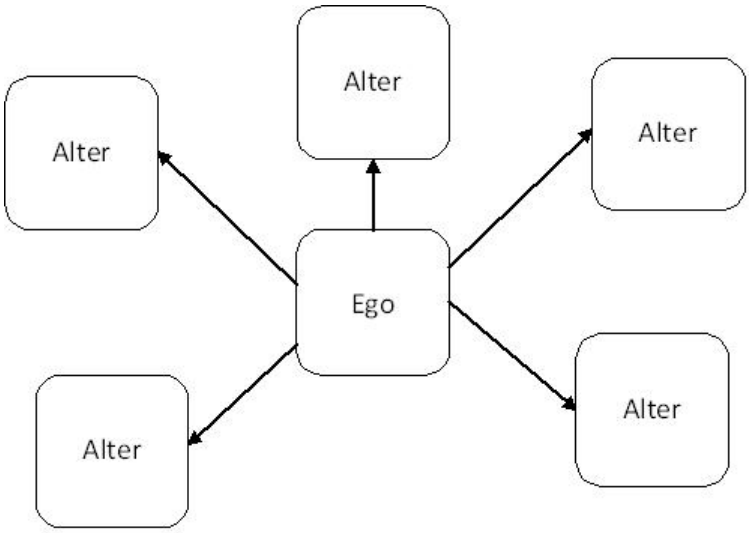

Figure 1.1: An Egocentric network with the "Ego" at the centre and five "Alters" forming dyadic relationships.

the dyadic relationships formed from the ego would be the approach of an egocentric network (Friedman and Aral, 2001). Egocentric risk networks ask the ego to name people in his/her risk network, with risk defined by either sexual relationship and/or injecting drugs. A sex trade worker naming five people who he/she has had unprotected sex with, is an example of a sexual risk network. This group of people would then form an egocentric sexual risk network, with the link, (risk) being unprotected sex between the ego and each alter (Figure 1.1). The link or arrow in the figure is called a dyadic relationship. This represents the bond between the two people. The diagram of an egocentric network would represent ego at the centre forming a dyadic relationship to each person named, the Alters.

The second type of social network is known as Sociocentric (Figure 1.2). Compared to egocentric, sociocentric risk networks deal with larger groupings of PWUID and not just the single Person Who Uses Injection Drugs. The links between these people form lines which create "sociograms" developed by Sociologists (Scott, 2000). The link represents high-risk behaviour like syringe sharing. Sociocentric risk networks delineate the paths through which viruses, such as Human Immunodeficiency Virus, (HIV) can be passed.

In a risk network, along with other types of networks, many different factors can pass through these links. Information is one such factor that can be passed amongst people in a network. One example comes from Central and Eastern Europe

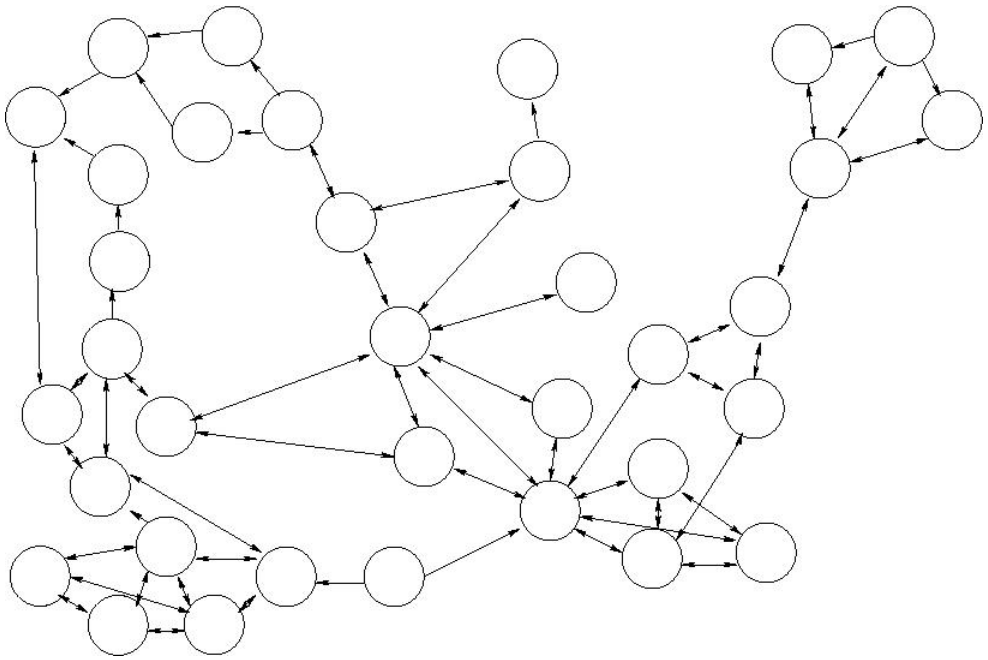

Figure 1.2: An example of a Sociocentric network with each circle representing a person and the arrow representing the relationship between those two people.

where network leaders were trained to carry out HIV prevention messages into regular everyday communication with friends. Network members were identified as leaders from the sociometric interviews and baseline risk assessments. Studies showed that at three and twelve month follow-ups, network members had a sharp reduction in behavioural risk (Amirkhanian et al., 2003). If people pass along information about harm reduction this would help to reduce the overall risk in the network. An example of beneficial information that is passed through a network is the use of condoms. If a sex trade worker were to use condoms he or she would greatly reduce the risk of their sexual risk network. If a network of people who use injection drugs were to only use clean needles this would greatly reduce the risk in their network (Wood et al., 2002).

Along with beneficial information being passed through these links, harmful information can also move throughout the network. PWUID can pass on high-risk sharing behaviours. An example of harmful information for people who use injection drugs is going to a shooting gallery. A shooting gallery is a place for PWUID to go and rent syringes and needles for drugs. Shooting galleries have been associated with a higher increase in blood borne viruses (Chitwood et al., 1990).

1.2 Peer Networks - Respondent Driven Sampling

One problem with facility and survey-based data collection is that they are not able to reach hidden populations (Magnani et al., 2005). Hidden populations can be sub-populations like injection drug users who partake in an illegal activity. This can make them harder to reach because of criminal sanctions. However, it is very important for HIV research and intervention to be able to reach these people, who constitute a high-risk group. To combat the inefficiencies involved with facility and survey-based sampling there are several different types of sampling methods that can be used. One such method, that will be discussed here, is respondent driven sampling (RDS).

Respondent driven sampling is a spawn of chain-referral sampling and it was formed to reduce the bias of such things like masking and volunteerism (Heckathorn et al., 2002). Masking is the bias associated with an interview question that deals with a sensitive subject. People are less likely to correctly answer sensitive questions. Volunteerism is the opposite, where people are more likely to answer certain questions. This also produces bias, as those questions will be over represented. Respondent driven sampling involves a system where a subject is given a reward for participating in the sample but also given a reward for referring or recruiting others to participate. This is different from snowball sampling, where one is only rewarded for participating in the survey (Heckathorn, 1997). This two stage system is a way to reach hard to sample sub-populations that are important in HIV intervention, such as People Who Use Injection Drugs. Not only are subjects being rewarded with a cash incentive they are also being rewarded with a chance to help peers from a deadly epidemic (Heckathorn, 1997).

Heckathorn et al. (2002) interviewed 386 injection drug users from Meriden, Connecticut via respondent driven sampling. They interviewed PWUID who were over the age of eighteen and had injected drugs in the past 30 days. This study had two phases in which "steering incentives", a supplementary reward for recruiting a certain category of subjects, were used in one stage and the analysis on this was computed. The point of the study was to see if steering incentives were beneficial in recruiting a specific target population in respondent driven sampling. The incentive in this case was a ten dollar bonus given to interviewees if they recruited an injection drug user between the ages of 18-25. The respondent was then rewarded an extra ten dollars for doing the survey. The survey found that steering incentives were indeed significant in increasing the number of 18-25 year olds contacted. Before steering incentives were

used there were 45 out of 196 that fell in to this age category. After steering incentives the number of 18-25 year olds increased to 74 out of 190.

Stormer et al. (2006) performed a study in St. Petersburg, Russia and Tirana, Albania where they interviewed participants that were over the age of 15, who had recreationally injected drugs in the past six months. Interviews in both cities were done face-to-face over a 7-8 week period. In Tirana 15 seeds were selected and in return recruited 210 injection drug users (IDU's) during July and August of 2005. In St. Petersburg a total of 13 seeds were selected and they recruited 200 IDU's during the same time span. The goal of this study was to compare implementation of RDS, network analysis, and the properties of recruitment. Results showed that respondent driven sampling was an effective way of getting at a hard-to-reach population. RDS also showed that the networks were comprised differently between the two cities. Tirana had less women IDU's than men, in comparison to St. Petersburg where there were more women.

1.3 Women

Women who are injection drug users have a greater risk of obtaining HIV and other infectious diseases, such as hepatitis C. Women Who Use Injection Drugs (WWUID), are more likely to have a male sex partner who also uses drugs (Miller and Neaigus, 2001). This relationship creates a greater risk for disease in women. Drug users may believe that sharing needles with a sex partner is more acceptable than sharing needles with a stranger (Davey-Rothwell and Latkin, 2007). Studies have shown that needle sharing happens to occur in sexual relationships where there is some emotional attachment (Unger et al., 2006). Relationship dynamics can play a major role in needle sharing behaviour. A partner's refusal to share needles could be perceived as a form of mistrust. WWUID, that are dependent on their partner for support such as money, housing or protection, might feel less empowerment to voice their opinion on the use of clean needles (Unger et al., 2006).

Violence is a severe problem among WWUID. Women who have a history of abuse and sexual assault have a greater chance of drug dependence or abuse compared to women with no such history (Whynot, 1998). Young homeless women who are surrounded by all types of violence are directed into entering a relationship with older men. Relationships such as these are often terrible for both people; lopsided power control and abusiveness make females very vulnerable (Bourgois et al., 2004).

Different factors that increase HIV risk amongst women may occur at different causal levels (Miller and Neaigus, 2001). If we understand these levels we can then develop the appropriate form of intervention to help curb disease transmission. One of these levels that is of importance is the network.

Freeman, Rodriguez, and French (1994) analyzed a survey in New Jersey, the Paterson Health Behavior Project (PHBP). The PHBP was a project that recruited injection drug users and their sex partners. They were interviewed on their sexual and needle risk behaviour related to HIV. After the study subjects were offered pre/post test counseling along with an HIV test and an assortment of intervention programs. The goal of this project was to assess any differences of HIV predictors associated with gender. Results showed that female injection drug users living in Paterson could be at greater risk of acquiring HIV because of their involvement with a sex partner who uses injection drugs. Females were significantly more likely than males to have been injected by their male sex partner after he had already injected himself. This act greatly increases the chances of contracting HIV if there isn't any rigorous cleansing of the needle.

1.4 Norms

Social networks can play a role, either indirectly or directly, in influencing our behavior. Norms are an acceptable or a typical behaviour used amongst a social network (Davey-Rothwell and Latkin, 2007). This behaviour could be an idea or belief that is common practice throughout the network. According to Davey-Rothwell and Latkin (2007) by the observance of others' behaviour, norms start to form what is appropriate. People will compare their own behaviour to that of someone else and then pattern their actions accordingly. Perceived norms are comprised of two corresponding ideas. An individuals' view of the actions of others' behaviours are known as descriptive norms. The other, injunctive norms, is the pressure we feel from the behaviours people approve or disapprove of (Davey-Rothwell and Latkin, 2007). It is important to try and understand perceived norms to further understand HIV risk behaviour. If people are more influenced by others with similar attitudes and behaviour then PWUID are at a greater risk if they are in a relationship with an injection drug user partner compared to that of a non injection drug user partner (Davey-Rothwell and Latkin, 2007). Descriptive norms have been found to be more significant when dealing with observable injection behaviour (Davey-Rothwell and Latkin, 2007).

In Baltimore, Maryland, a social network-based HIV prevention intervention called "STEP into action" (STEP) was used to collect cross-sectional data. The data came from participants who were recruited through street outreach. Participants were eligible if they were at least 18, lived in Baltimore, and had injected in the past six months. Furthermore, they couldn't have participated in a previous HIV or network study within the last year and had to be willing to speak to a fellow network member about HIV risk reduction. The point of this study was to assess the link between HIV-based communication and perceived norms among PWUID from an intervention at the network level. The results were able to show that there is a link between HIV-based communication and perceived norms. Less HIV-based communication in a network promotes risky behaviour (Davey-Rothwell and Latkin, 2007).

Between May 2002 and January 2004 participants were recruited in Baltimore, Chicago, Los Angeles, New York, and Seattle. Through street outreach and respondent driven sampling 3285 participants were recruited in this study. The goal of the study was to examine the risk factors associated with PWUID and receptive syringe sharing (RSS). Results showed that peer norms (perceived risks and peer influences) for syringe sharing along with type of injection partner were associated with RSS (Hagan et al., 2007).

Social network analysis provides a unique approach to the study of people who use injection drugs. It is a very efficient tool for analyzing the many different factors that flow throughout the network, such as gender and norms. Instead of dealing with a person and trying to learn about his or her risk reduction practices we can instead focus on the whole community or sub-population. This allows us to gather a wide variety of information to help provide harm reduction practices in certain sub-populations. In order to sample hard to reach sub-populations, response driven sampling has shown to be an effective way of tackling this problem. RDS is important when dealing with PWUID and is a useful way of reaching this hidden sub-population. From this we can conclude that social network analysis is a viable option in the study of PWUID.

1.5 Discussion

Now that I have set up the basic principles of social network analysis, I would like to move forward and discuss how to analyze different types of data. Dyadic data is one type of data commonly seen in network analysis. I will focus on the problems that

arise with the analysis of dyadic data. Also, I will build suitable models for this type of data and compare them using frequentist and Bayesian methods.

Chapter 2

Random Effects Models

2.1 Generalized Linear Models

Generalized linear models (GLM's) are used when dealing with response variables that are continuous and not necessarily Gaussian distributed, or discrete. They are a general form of regression models that may use fixed and/or random effects. Some of the most commonly used GLM's are log-linear, logistic, and linear regression models. GLM's have three components that must be specified. First we must confirm that the response y belongs to the exponential family. To do so its density must be written in the following format

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \tag{2.1}$$

where a and s are known functions of y, and b and t are known functions of θ (Dobson, 2002), where θ is the family parameter. There are many well known distributions that are in the exponential family of distributions. Some such commonly known distributions are the Poisson, Binomial, and Gaussian. In the simple case, a distribution belonging to the exponential family can be arranged to be in canonical form as follows

$$exp\{a(y)b(\theta) + c(\theta) + d(y)\} \tag{2.2}$$

where a, b are the functions stated above, c, d are known functions and $a(y) = y$ (Dobson, 2009). Next, we can specify that the linear predictor or linear function of the predictors, η_i, is of the form

$$\eta_i = x_i^{'}\boldsymbol{\beta} \tag{2.3}$$

The linear predictor, η_i, is a linear combination of independent covariates, x_i and the fixed parameter, $\boldsymbol{\beta}$ where $i = 1..n$. Finally, a link function, g, such that it relates the mean, μ_i, to the linear predictor, η_i, is chosen.

$$g(\mu_i) = \eta_i \tag{2.4}$$

Some examples of commonly used link functions are provided in Table 2.1.

Table 2.1: Examples of some common link functions.

Distribution	Link Function
Binomial	$ln\left(\frac{\mu}{1-\mu}\right)$
Poisson	$ln(\mu)$
Gaussian	μ
Exponential	μ^{-1}

Example 2.1. *An example of a simple glm with response, y, following a Binomial distribution with a single covariate, x_i:*

$$y_i \sim Bin(n_i, n_i) \tag{2.5}$$

$$g(p) = ln\left(\frac{p}{1-p}\right) = \boldsymbol{x_i'\beta} \tag{2.6}$$

Therefore, using the logit link function, the distribution for $y = (y_i, .., y_n)$, y_i independent, would take the form

$$p(y|\beta) = \prod_{i=1}^{N} \binom{n_i}{y_i} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+e^{\eta_i}}\right)^{n_i - y_i} \tag{2.7}$$

2.2 Generalized Linear Mixed Models

Independence of the observations is one of the assumptions made with generalized linear models. However, many types of data are correlated by the nature of their design. One example is clustered data where subjects are not sampled independently, but instead nested within a larger group. For example, this happens when you collect data on individual students within schools (Figure 2.1), or even students within schools, within school districts.

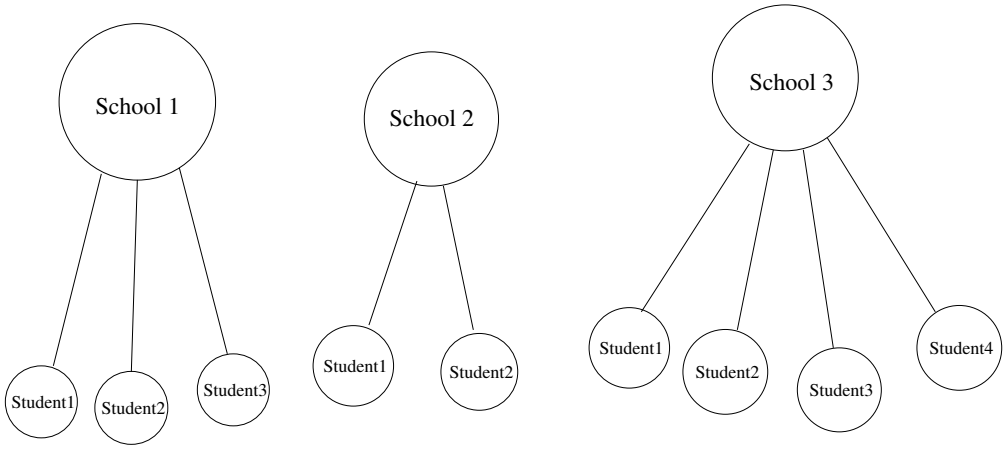

Figure 2.1: A diagram of a two-level model of students within schools.

We can have several different characteristics of students but if we ignore the characteristics of schools we would not be taking into consideration the significance of group effects (Goldstein, 2003). Designs such as these are called multilevel, hierarchical, nested, or mixed effects. Mixed effects modeling has many different uses in various disciplines.

In the social sciences, in particular education research, these designs are known as multilevel models (Dedrick et al., 2009). Hierarchical models are another way of describing multilevel models but are used within a Bayesian framework. In ecology and evolution, generalized linear mixed models are more commonly used (Bolker et al., 2008). The common thread is the analysis of correlated data, collected at multiple levels.

An extension to GLM's to allow for multilevel data are Generalized Linear Mixed Models (GLMMs) which use both fixed and random effects. There are several reasons to include random effects. One is to incorporate the variation amongst individuals that happens over repeated measures (Bolker et al., 2008). Another reason would be due to sample clustering where the same measurement is repeated causing a cluster of measurements (Agresti et al., 2000).

Pseudoreplication is the result of treating replicates as being independent when in fact they happen to have some level of dependence (Hulbert, 1984). Using the previous student and school example, if we were to assume that each student is independent of one another this would result in pseudoreplication. Since students are grouped according to their school, this creates some dependence. The students of a

particular school are in the same classes and being taught the same material through the same teachers. It would be wrong to assume each student is independent from one another.

The process of setting up a generalized linear mixed model is similar to that of a GLM. We must specify that the conditional distribution of the response given the random effects has a density in the form of the exponential family.

Next we specify the linear predictor. The linear predictor, η_i, is a linear combination of the independent covariates, x_i and z_i, the fixed parameters, $\boldsymbol{\beta}$, and the random parameters, $\boldsymbol{u}$. For example,

$$\eta_i = x_i^{'}\boldsymbol{\beta} + z_i^{'}\boldsymbol{u} \tag{2.8}$$

We must also specify a link function, which relates the linear predictor to the mean of the specified distribution, μ_i

$$g(\mu_i) = \eta_i \tag{2.9}$$

Finally, we specify the structure of the random effects and their relation to the linear predictor

$$\boldsymbol{u} \sim Norm(0, \tau) \tag{2.10}$$

where τ is the precision.

Example 2.2. *An extension to Example 2.1 by adding a random effect, b_i to the intercept; b_i represents the dependence of subject i on its covariates.*

$$y_i \sim Bin(n_i, p_i) \tag{2.11}$$

$$logit(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + b_i \tag{2.12}$$

2.3 Multilevel Modeling

A two-level model would have the first level observations which are the subjects, nested within the second level, which are the clusters (Figure 2.1). When building a multilevel model it is important to start with the first level, before moving on to examine the predictors in the second level. This is done so that the need for a multilevel model is attained. If you build the first level and realize there is no need for a multilevel model then there is no need to look at level 2. If, by examining the

data and the predictors, we find there is sufficient need for a hierarchical model, then we start by building the second level model with an intercept-as-outcomes model. An intercept-as-outcomes model is a model with a random intercept that varies according to the level 2 predictors. Next, the data should be examined for the need of a slopes-as-outcomes model. A slopes-as-outcomes model is used when there is need for a random slope component that varies according to the level 2 predictors (Luke, 2004).

The following is a simple linear regression model with response variables $y_1, y_2, .., y_n$, covariates $x_1, x_2, .., x_n$, intercept α, slope β, and normally distributed independent errors ϵ_i:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{2.13}$$

To incorporate random effects we substitute β_{0j} for our intercept α and β_{1j} for our slope β where i associated with level 1 is $i = 1, 2, .., n_j$ and j associated with level 2 is $j = 1, 2, .., J$, which gives:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij} \tag{2.14}$$

where β_{0j} and β_{1j} are equal to:

$$\beta_{0j} = \beta_0 + u_{0j} \tag{2.15}$$

$$\beta_{1j} = \beta_1 + u_{1j} \tag{2.16}$$

where u_{0j} is the random component of the intercept and u_{1j} is the random component of the slope. If we substitute 2.15 and 2.16 into 2.14 we obtain

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0j} + u_{1j} x_{ij} + \epsilon_{ij}) \tag{2.17}$$

Here y_{ij} is modeled as the fixed effects β_0 and β_1, and the random effects u_{0j}, u_{1j}, and ϵ_{ij}.

We can extend the two-level model to incorporate another level of depth. To visualize this we can refer to the example of childhood behaviour problems (Romano et al., 2005). Here, children were nested within families, nested within neighbourhoods.

Algebraically, a three level model would look as follows:

$$y_{ijk} = \beta_{0jk} + \beta_{1jk}x_{ijk} + \epsilon_{ijk} \tag{2.18}$$

where i associated with level 1 is $i = 1, 2, .., n_{jk}$, j associated with level 2 is $j = 1, 2, .., J_k$, k associated with level 3 is $k = 1, 2, .., K$, and β_{0jk} and β_{1jk} are equal to:

$$\beta_{0jk} = \beta_0 + v_{0k} + u_{0jk} \tag{2.19}$$

$$\beta_{1jk} = \beta_1 + v_{1k} + u_{1jk} \tag{2.20}$$

where v_{0k} and v_{1k} are the random components for level 3, and u_{0jk} and u_{1jk} are the random components for level 2. If we substitute 2.19 and 2.20 into 2.18 we get

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + (v_{0k} + u_{0jk} + v_{1k}x_{ijk} + u_{1jk}x_{ijk} + \epsilon_{ijk}) \tag{2.21}$$

This is a random slope and intercept model with random effects v_{0k}, u_{0jk}, v_{1k}, u_{1jk}, and ϵ_{ijk} and fixed effects β_0 and β_1.

2.4 Model Selection

Model selection is a very important part of analysis. Since our models were fitted using maximum likelihood and Bayesian methods, we need to look at two separate model selection technique: Akaike's Information Criterion (AIC) for our classical models and Deviance Information Criterion (DIC) for our Bayesian models.

2.4.1 AIC

With linear models, model selection can be done using Akaike's Information Criterion (AIC) or for small samples a corrected AIC_c is used (Burnham and Anderson, 2002). These are defined as

$$AIC = -2\log(L(\hat{\theta}|y)) + 2k \tag{2.22}$$

$$AIC_c = -2\log(L(\hat{\theta}|y)) + 2k\frac{k+1}{n-k-1} \tag{2.23}$$

where n is the sample size, k is the number of estimable parameters, and $L(\hat{\theta}|y)$ is the likelihood evaluated at the maximum likelihood estimate $\hat{\theta}$. When comparing models, one constructs a set of candidate models and compares AIC values with that of the minimum AIC in the candidate set; these are defined as AIC differences ($\Delta_i = AIC_i - AIC_{min}$).

To evaluate models further we can assess Akaike weights (w_i) which can be interpreted as the relative likelihood of a model given the data and the model set.

$$w_i = \frac{\exp(-1/2\Delta_i)}{\sum_{r=1}^{R} \Delta_r} \tag{2.24}$$

where R is the number of models in the candidate set. A particular model weight can be thought of as the evidence in favour of model i (*i.e.* the larger the model weight, the more evidence in favour of model i within the set of candidate models).

As with simple linear regression models, model selection for multilevel models can be done using Akaike's Information Criterion (AIC). The AIC we used for model selection is the marginal AIC. This was used since we are not interested in the random effects as a parameter. We can ignore it when making inference. The formula can be seen below (Vaida and Blanchard, 2005):

$$mAIC = -2 \log g(y|\hat{\theta}) + 2K \tag{2.25}$$

where K is the number of parameters of the estimate $\hat{\theta}$ and $g()$ is the marginal observed likelihood.

2.4.2 DIC

The DIC is a Bayesian alternative to using AIC when dealing with hierarchical models (Spiegelhalter et al., 2002). To get our DIC values we use a program called WinBUGS (Lunn et al. 2000). WinBUGS is statistical software that allows us to run models and analyses using a Bayesian framework. WinBUGS has a DIC calculator built in and was used for our model selection. The DIC formula can be seen as follows:

$$DIC = D(\hat{\theta}) + 2p_D \tag{2.26}$$

where $D(\hat{\theta})$ is the measure of fit for the data -a point estimate of the deviance evaluated at the posterior mean- and is calculated as $D(\hat{\theta}) = -2log(p(y|\bar{\hat{\theta}}))$; $\bar{\hat{\theta}}$ is the mean

of the posterior samples of θ. The effective number of parameters ($p_D = \bar{D} - D(\hat{\theta})$) is defined as the posterior mean of the deviance minus the deviance of the posterior mean. Like Δ_i, ΔDIC_i is defined as DIC of the model of interest minus the minimum DIC value in the candidate set.

2.5 Inference: Maximum Likelihood

Maximum likelihood (ML) inference requires the development of the likelihood. We start by looking at the logistic mixed model. Consider the response, y_{ij}, to be either a 0 or a 1 thus modeled using a Bernoulli distribution. We will assume the random effects (b_i) are approximately $i.i.d$ normal with mean 0 and standard deviation σ.

$$y_{ij}|b_i \sim Bern(p_{ij}) \tag{2.27}$$

We can see, when using a mixed model, that the response is now conditional on the random effects, b_i. Using the logit link brings us to the following,

$$log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + b_i \tag{2.28}$$

where

$$b_i \sim Norm(0, \sigma^2) \tag{2.29}$$

Setting $log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij}$ we can rearrange and solve for p_{ij}. Therefore,

$$p_{ij} = \frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}} = \frac{e^{\beta_0+\beta_1 x_{ij}+b_i}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.30}$$

Having set up the model we can now begin to find the likelihood. First,

$$L(\boldsymbol{\beta}, \sigma^2) \propto Pr(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \tag{2.31}$$

We cannot write out $Pr(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$ because the exact form is not known, we do however have some information that will help us solve it. We know that $y_{ij}|b_i \sim Bern(p_{ij})$ and we also know $b_i \sim Norm(0, \sigma^2)$. Therefore we can integrate out the random effects as follows,

$$Pr(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} Pr(\boldsymbol{y}, \boldsymbol{b}|\boldsymbol{\beta}, \sigma^2) d\boldsymbol{b} \tag{2.32}$$

where $\boldsymbol{b} = (b_1, .., b_N)^t$. Now the joint probability $Pr(\boldsymbol{y}, \boldsymbol{b}|\boldsymbol{\beta}, \sigma^2)$ is just the probability of $\boldsymbol{y}$ given the random effects multiplied by the probability of those random effects.

$$Pr(\boldsymbol{y}, \boldsymbol{b}) = Pr(\boldsymbol{y}|\boldsymbol{b})Pr(\boldsymbol{b}) \tag{2.33}$$

where,

$$Pr(\boldsymbol{y}|\boldsymbol{b}) = \prod_{i=1}^{N}\prod_{j=1}^{N_i} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \tag{2.34}$$

and

$$Pr(\boldsymbol{b}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{b_i^2}{2\sigma^2}} \tag{2.35}$$

Substituting in equation 2.25 we get,

$$Pr(\boldsymbol{y}|\boldsymbol{b}) = \prod_{i=1}^{N}\prod_{j=1}^{N_i} (\frac{e^{\beta_0+\beta_1 x_{ij}+b_i}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}})^{y_{ij}} (1 - \frac{e^{\beta_0+\beta_1 x_{ij}+b_i}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}})^{1-y_{ij}} \tag{2.36}$$

$$Pr(\boldsymbol{y}|\boldsymbol{b}) = \prod_{i=1}^{N}\prod_{j=1}^{N_i} \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.37}$$

$$Pr(\boldsymbol{y}, \boldsymbol{b}) = \prod_{i=1}^{N}\prod_{j=1}^{N_i} \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{b_i^2}{2\sigma^2}} \tag{2.38}$$

$$Pr(\boldsymbol{y}, \boldsymbol{b}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{b_i^2}{2\sigma^2}} \prod_{j=1}^{N_i} \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.39}$$

$$Pr(\boldsymbol{y}) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{b_i^2}{2\sigma^2}} \prod_{j=1}^{N_i} \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} db_1..db_N \tag{2.40}$$

$$Pr(\boldsymbol{y}) = \prod_{i=1}^{N} \left[\int_{-\infty}^{\infty} (2\pi)^{-\frac{1}{2}} \sigma^{-1} e^{-\frac{b_i^2}{2\sigma^2}} \prod_{j=1}^{N_i} \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} db_i \right] \tag{2.41}$$

2.6 Parameter Estimation

There are several methods for the estimation of parameters in multilevel models. Some approximate methods are maximum likelihood (ML), penalized quasi-likelihood

(PQL), generalized estimating equations (GEE), and Bayesian methods. These are called approximate methods since the likelihood, which sometimes can be difficult to calculate, is estimated. There are also some sampling type methods of calculating these estimates such as Markov chain Monte Carlo (MCMC) and a Monte Carlo EM algorithm (MCEM). The first method, ML, is the most commonly used estimation technique. It works very well in models that are not complex. The likelihood is as follows,

$$\boldsymbol{L} = \int \prod_i f_{y_i|\underline{u}}(y_i|\underline{u}) f_{\underline{u}}(\underline{u}) d\underline{u} \tag{2.42}$$

where y_i is the response, $\underline{\mathrm{u}}$ is the random parameters, $f_{y_i|\underline{u}}$ is the distribution with respect to the response, and $f_{\underline{u}}$ is the distribution with respect to the random parameters.

The likelihood can be calculated using numerical methods, such as Newton-Raphson an optimization technique, when there are few random effects (McCulloch and Searle, 2001). However, when models of higher order structure are used, integration by numerical analysis will break down. With increasing higher order structures comes integrals increasingly difficult to numerically integrate.

Penalized quasi-likelihood (PQL) uses a Laplace approximation to the quasi-likelihood to produce estimates. Quasi-likelihood was defined by Wedderburn (1974) as a way of defining the likelihood without having to know the distribution of the observations. Instead, we only need to specify a relationship between the mean and variance. Since there are quite a few approximations carried out when finding PQL estimators, bias will be present in the estimators (de Leeuw and Meijer, 2008).

There is no closed form for the function $Pr(\boldsymbol{y})$. Therefore, we can turn to approximation methods such as Newton's Method as an alternative. Newton's Method is a recursive procedure that starts with an initial value and computes the next value from the initial value. Seen here,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{2.43}$$

where $f(x_0)$ is the function to be approximated, $f'(x_0)$ is it's derivative, and x_0 is the first guess for a solution. The above equation is then repeated until an approximation is met.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.44}$$

2.7 Inference: Bayesian methods

In order to do inference using Bayesian methods we need to use what is called a posterior density ($p(\theta|y)$). To find the posterior we first use the prior information we have on our parameter θ ($p(\theta)$) and multiply it by the sampling distribution of our data ($p(y|\theta)$). This forms the joint probability distribution which can be seen below.

$$p(\theta, y) = p(\theta)p(y|\theta) \tag{2.45}$$

Now using the joint probability distribution and the marginal distribution of y we can solve for the posterior.

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{2.46}$$

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{2.47}$$

2.8 Markov chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a tool used to sample from probability distributions. This is useful in Bayesian statistics where we need to sample from a probability distribution that contains a multi-dimensional integral. The first step in MCMC is to construct a Markov chain that will eventually converge to the desired distribution of interest, the posterior. After running the simulation long enough we will eventually be able to sample from the desired posterior distribution. We can control the number of "steps" of the Markov chain and by doing this the precision of the samples. One thing to note is that in a Markov chain each step depends on the previous thus making the samples drawn dependent samples.

2.8.1 Gibbs sampling

One MCMC algorithm of particular note is Gibbs sampling. Gibbs sampling is used when the joint distribution is not known in its exact form or when dealing with a high degree could be difficult to sample. How Gibbs sampling helps is that all we need to know is the full conditionals of each variable in the joint distribution. The full conditional of a variable is the probability of that variable given all other variables in

that model. These full conditionals will be easier to sample from then the posterior and the result produces a Markov chain from the iterative estimates. The stationary distribution of this Markov chain will be our desired joint distribution, the posterior distribution.

In equation 2.39 $p(y|\theta)$ is the likelihood, $p(\theta)$ is the priors on β, b_i, and σ^2 and these can be seen below.

$$p(y|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{N_i} \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.48}$$

$$b_i|\sigma^2 \propto N(0,\sigma^2) \tag{2.49}$$

$$\sigma^2 \propto Inv-Gamma(.001,.001) \tag{2.50}$$

$$\beta \propto N(0,10^6) \tag{2.51}$$

We are using non informative priors for β and σ^2. Subbing equation 2.40, 2.41, 2.42, and 2.36 into equation 2.39 we get,

$$p(\beta,\sigma^2,b_i|y) \propto Pr(\beta)P(\sigma^2)P(b_i)P(y) \tag{2.52}$$

$$p(\beta,\sigma^2,b_i|y) \propto \prod_{i=1}^{N}\prod_{j=1}^{N_i} \sigma^{-3.002} exp(\frac{-\beta^2}{(2)(10^6)} - \frac{.001}{\sigma^2} - \frac{b_i^2}{2\sigma^2}) \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.53}$$

In order to sample we need to derive the full conditionals for each parameter. Therefore, the full conditionals for β, σ^2, and b_i are

$$P(\beta_0|y,\sigma^2,b_i) \propto exp(\frac{-\beta_0^2}{2*10^6}) \prod_{i=1}^{N}\prod_{j=1}^{N_i} \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.54}$$

$$P(\beta_1|y,\sigma^2,b_i) \propto exp(\frac{-\beta_1^2}{2*10^6}) \prod_{i=1}^{N}\prod_{j=1}^{N_i} \frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.55}$$

$$P(b_i|\sigma^2,\beta,y) \propto \frac{1}{\sigma}\prod_{i=1}^{N}\prod_{j=1}^{N_i} exp(\frac{-b_i^2}{2*\sigma^2})\frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.56}$$

$$P(\sigma^2|\beta,y,b_i) \propto (\sigma^2)^{-(1.001)} exp(\frac{-0.001}{\sigma^2})\prod_{i=1}^{N}\prod_{j=1}^{N_i}\frac{(e^{\beta_0+\beta_1 x_{ij}+b_i})^{y_{ij}}}{1+e^{\beta_0+\beta_1 x_{ij}+b_i}} \tag{2.57}$$

2.9 Hypothesis testing

For generalized linear mixed models, hypothesis testing is done by using a Wald test of the hypothesis of no effect of overdispersion. When there is no overdispersion, Wald Z and χ^2 tests can be used. If there is overdispersion in the data Wald t and F tests can be used to deal with the uncertainty in the estimates (Bolker et al., 2008). With nested models, the likelihood ratio test statistic can be used although it is not recommended when sample sizes are small (Bolker et al., 2008). It is computed as (Λ), where (Λ), the deviance is a difference in likelihoods. It is an asymptotic approximation since the exact likelihood in most cases cannot be computed. One thing to note is that the likelihood ratio statistic does not follow a simple χ^2 distribution (McCulloch, 2003), rather it follows a 50:50 mixture of a χ_1^2 and 0 when testing the hypothesis $H_o : \sigma^2 = 0$, where σ^2 is a single variance component. With large sample sizes, likelihood ratio tests are preferred over Walk tests for inference on random effects (Bolker et al., 2008).

2.10 Assessing model fit

Assessing model fit is somewhat of a problem with multilevel models. The usual goodness-of-fit tests for linear regression don't simply transfer across to multilevel models. The debate in the literature on how to assess model fit is ongoing and there is no consensus on the best procedure.

2.11 Diagnostics

Model checking via residuals is done at each level. In a two-level model there is a set of residuals for both levels. There are two competing methods of model checking. The first is to start with residual checking of the first level proceeding to the next

higher level. The second, is to start at the highest level working downwards to the first level. The literature is split on which way residual analysis should be done. According to Snijders and Berkhof (2008) the best approach is to start at level one and work downwards. This way when level one's residuals are analyzed, the next levels have no confounding effects.

One way to see which fixed effects are significant is to add additional fixed effects to a specified model then test the significance of the two nested models (Pan and Lin, 2005).

2.12 Centering

In multilevel models one question that arises is whether or not the predictor variables should be centered. The answer to that question largely depends on what is the predictor variable. Centering variables, in a multilevel model, removes the collinearity between the predictors in the model; with collinearity the estimates might be poorly estimated (Paccagnella, 2006). However, we need to be careful when centering as it can change how we interpret an estimate. If we are to examine centering in linear regression, we would see that by centering we are offsetting the intercept. To see this we take a linear regression model,

$$y_i = \beta_0 + \beta_1 x_i \tag{2.58}$$

now we center by subtracting the mean, $\bar{x}$ from each covariate,

$$\begin{aligned} y_i &= \beta_0 + \beta_1(x_i - \bar{x}) \\ &= \beta_0 + \beta_1 x_i - \beta_1 \bar{x} \end{aligned} \tag{2.59}$$

Here $\beta_1 \bar{x}$ is just a constant, so it can be grouped with the intercept β_0. Which yields,

$$y_i = \beta_0^* + \beta_1 x_i \tag{2.60}$$

where

$$\beta_0^* = \beta_0 - \beta_1 \bar{x} \tag{2.61}$$

The interpretation of these parameters is as follows, before centering, the intercept,

β_0, is the expected outcome when all the explanatory variables have a zero value. After centering, the intercept, β_0^*, becomes the expected outcome when all the explanatory variables are equal to the mean. This is called grand mean centering and is the same for linear regression and multilevel models (Paccagnella, 2006).

Another way of centering is called group mean centering. This is done by centering the intercept around each group's mean. If we are to refer back to the student within schools example, we would center all the students in school_j with the mean of all the students in school_j. Since we are subtracting different values the intercepts will be harder to interpret. We would now interpret the intercept as being the expected outcome of group_j when the explanatory variables of group_j are equal to the mean of group_j.

One of the main benefits of multilevel models is the ability to properly analyze data of a hierarchical order. They work well in analyzing the relationships of groups when applying that to individuals or vice versa.

One of the problems with multilevel models is in how they deal with missing values. According to Collins et al. (2001) missing values may introduce some bias in the parameter estimates. On top of that, depending on the complexity of the data it may be hard to accurately approximate these missing values using classical analysis.

Chapter 3

Examples of Multi-level Model Fitting

In this chapter we study a real data set on seed germination as well as a simulated data set that is similar to what we would find with our network data in chapter 4.

3.1 Seed Germination Example

This is an example of a 2x2 factorial layout, focusing on the number of seeds germinated. Factors are type of seed and root extract. Analysis was done in R using the glmmML package for maximum likelihood estimates and done in WinBUGS for Gibbs sampling estimates.

Data for Example 2.3 are from the 1978 paper by Crowder and can be seen in Table 3.1.

Table 3.1: Bean and cucumber root extract data for seeds *O. aegyptiaca* 75 and *O. aegyptiaca* 73.

O. aegyptiaca 75						O. aegyptiaca 73					
Bean			Cucumber			Bean			Cucumber		
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
10	39	0.26	5	6	0.83	8	16	0.50	3	12	0.25
23	62	0.37	53	74	0.72	10	30	0.33	22	41	0.54
23	81	0.28	53	72	0.76	8	28	0.29	15	30	0.50
26	51	0.51	32	51	0.63	23	45	0.51	32	51	0.63
17	39	0.44	46	79	0.58	0	4	0.00	3	7	0.43
			10	13	0.77						

The model used in WinBUGS is as follows,

$$r_i \sim Bin(p_i, n_i) \tag{3.1}$$

$$logit(p_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i \tag{3.2}$$

$$b_i \sim Norm(0, \tau) \tag{3.3}$$

where r_i is the number of seeds germinated on the ith plate, $i = 1, \cdots, N$, n_i is the total number of seeds on the ith plate, N is the total number of plates, x_{1i} is equal to 0 for seed type O. aegyptiaca 75 (and 1 otherwise) for the ith plate, x_{2i} is equal to 0 for the bean root extract (and 1 otherwise) for the ith plate, and b_i is the random effect for plate and is assumed to be Gaussian with mean zero and a uniform prior placed on the standard deviation. The alpha's were all given independent non informative priors. The results for the estimates can be seen in Table 3.2.

Table 3.2: Results from a Bayesian analysis of the data from Example 2.3. The posterior mean, standard deviation (SD) as well as quantiles from the posterior distribution are provided.

Parameter	Posterior Mean	SD	2.5%	Median	97.5%
α_0	-0.554	0.210	-0.977	-0.553	-0.139
α_1	0.060	0.358	-0.683	0.069	0.752
α_2	1.372	0.307	0.803	1.359	2.028
α_{12}	-0.841	0.500	-1.881	-0.821	0.126
τ	0.352	0.153	0.058	0.339	0.694

After fitting the model in WinBUGS, the same data were analyzed in R using the glmmML package; this provided maximum likelihood estimates. The results can be seen in Table 3.3.

Table 3.3: Results from the maximum likelihood analysis of the data from Example 2.3. The estimate, estimated standard error (SE), Z test statistic and p-value are provided for the hypothesis that the factor has no effect.

Parameter	Estimate	SE	Z	$\Pr(>\|z\|)$
α_0	-0.558	0.126	-4.429	9.46e-06
α_1	0.146	0.223	0.654	5.13e-01
α_2	1.318	0.178	7.428	1.10e-13
α_{12}	-0.778	0.306	-2.539	1.11e-02

We note that the ML estimates and posterior means are very close in value; however the error estimates obtained from the two methods differ with the Bayesian standard deviations being larger than the estimated standard errors. In terms of this data, using seed type O. aegyptiaca 73 increased the number of seeds germinated Similarly, using the cucumber extract also increased the number of seeds germinated. However; there was a negative interaction effect with the O. aegyptiaca 73 seed type and the cucumber extract such that when used together the number of seeds germinated was reduced.

3.2 Simulation Study

To study mixed effects models further in both the classical and Bayesian setting, we performed a simulation study in both R and WinBUGS. The data consist of 20 subjects answering 5 questions. Assuming complete independence there would be an N of size 100. However, there could be some dependence involved thus forming clusters. The cluster would be the answers the subject made. We could think of this as a multilevel model with the dependent clustered answers at one level and the subjects at another.

For each level there was one continuous and one categorical explanatory variables generated. The continuous generated variables were random standard normals. This could represent the age of the subject for level one (x_{1i}) and the age of the person the subject was answering questions about for level 2 (x_{2ij}). The categorical generated variables were random Bernoulli's ($p = 0.5$). These covariates could represent the sex of the subject at level one (z_{1i}) and the relationship status at level two (z_{2ij}). The responses were also generated as random binomials with parameter p where

$$logit(p) = \alpha_0 + \alpha_1 x_{1i} + \beta_1 z_{1ij} + \beta_2 z_{2ij} + b_i \tag{3.4}$$

where $i = 1, \cdots, 20$ and $j = 1, \cdots, 5$. Note that x_{2ij} was not actually used to generate the response and we should see it fall out of the model. The random effect was generated as a random normal variable ($\mu = 0, \sigma_b^2 = 2$). Table 3.4 shows the data generated for this simulation.

Table 3.4: Data generated from the model: $logit(p) = \alpha_0 + \alpha_1 x_{1i} + \beta_1 z_{1ij} + \beta_2 z_{2ij} + b_i$.

Individual	Y	x_1	z_1	z_2
1	0.00	1.33	1.00	1.00
1	1.00	1.33	0.00	0.00
1	0.00	1.33	0.00	1.00
1	1.00	1.33	1.00	0.00
1	1.00	1.33	0.00	1.00
2	1.00	-0.67	1.00	0.00
2	0.00	-0.67	1.00	1.00
2	1.00	-0.67	0.00	0.00
2	0.00	-0.67	1.00	0.00
2	0.00	-0.67	1.00	0.00
3	0.00	-0.97	0.00	1.00
3	0.00	-0.97	0.00	0.00
3	0.00	-0.97	1.00	0.00
3	1.00	-0.97	0.00	1.00
3	0.00	-0.97	0.00	0.00
4	1.00	0.28	1.00	1.00
4	1.00	0.28	1.00	1.00
4	1.00	0.28	0.00	1.00
4	1.00	0.28	1.00	1.00
4	1.00	0.28	0.00	0.00
5	1.00	-0.05	1.00	0.00
5	0.00	-0.05	1.00	1.00
5	1.00	-0.05	1.00	0.00
5	0.00	-0.05	0.00	1.00
5	1.00	-0.05	0.00	0.00

Individual	Y	x_1	z_1	z_2
6	0.00	-0.39	0.00	0.00
6	1.00	-0.39	1.00	0.00
6	1.00	-0.39	0.00	0.00
6	1.00	-0.39	0.00	0.00
6	1.00	-0.39	0.00	0.00
7	1.00	1.36	1.00	1.00
7	1.00	1.36	0.00	1.00
7	1.00	1.36	1.00	1.00
7	1.00	1.36	0.00	0.00
7	1.00	1.36	0.00	1.00
8	0.00	-2.26	0.00	0.00
8	0.00	-2.26	1.00	1.00
8	0.00	-2.26	0.00	1.00
8	0.00	-2.26	0.00	1.00
8	0.00	-2.26	0.00	0.00
9	0.00	-1.48	1.00	1.00
9	0.00	-1.48	0.00	0.00
9	0.00	-1.48	1.00	1.00
9	0.00	-1.48	0.00	1.00
9	0.00	-1.48	0.00	1.00
10	1.00	0.18	0.00	0.00
10	0.00	0.18	0.00	0.00
10	1.00	0.18	1.00	0.00
10	0.00	0.18	0.00	0.00
10	0.00	0.18	0.00	1.00
11	1.00	1.69	1.00	0.00
11	1.00	1.69	1.00	0.00
11	1.00	1.69	1.00	1.00
11	1.00	1.69	0.00	1.00
11	1.00	1.69	1.00	1.00

Individual	Y	x_1	z_1	z_2
12	1.00	-1.10	1.00	1.00
12	0.00	-1.10	1.00	1.00
12	0.00	-1.10	0.00	0.00
12	1.00	-1.10	1.00	1.00
12	1.00	-1.10	0.00	0.00
13	1.00	-0.87	1.00	0.00
13	0.00	-0.87	0.00	1.00
13	1.00	-0.87	1.00	0.00
13	1.00	-0.87	1.00	0.00
13	1.00	-0.87	1.00	1.00
14	1.00	0.65	0.00	0.00
14	1.00	0.65	0.00	1.00
14	1.00	0.65	1.00	1.00
14	1.00	0.65	1.00	1.00
14	1.00	0.65	1.00	0.00
15	1.00	0.73	1.00	0.00
15	1.00	0.73	0.00	0.00
15	1.00	0.73	0.00	1.00
15	1.00	0.73	1.00	1.00
15	1.00	0.73	1.00	1.00
16	1.00	1.02	1.00	0.00
16	1.00	1.02	0.00	0.00
16	1.00	1.02	1.00	0.00
16	1.00	1.02	0.00	1.00
16	1.00	1.02	1.00	1.00
17	0.00	-0.51	0.00	1.00
17	1.00	-0.51	1.00	0.00
17	1.00	-0.51	1.00	1.00
17	0.00	-0.51	0.00	0.00
17	0.00	-0.51	0.00	1.00
18	0.00	-0.93	0.00	0.00
18	0.00	-0.93	0.00	0.00
18	0.00	-0.93	0.00	0.00
18	0.00	-0.93	0.00	1.00
18	0.00	-0.93	0.00	0.00

Individual	Y	x_1	z_1	z_2
19	1.00	1.43	0.00	0.00
19	1.00	1.43	1.00	1.00
19	1.00	1.43	0.00	0.00
19	1.00	1.43	0.00	0.00
19	1.00	1.43	0.00	0.00
20	0.00	-1.54	0.00	1.00
20	1.00	-1.54	1.00	1.00
20	0.00	-1.54	0.00	0.00
20	0.00	-1.54	1.00	1.00
20	0.00	-1.54	0.00	0.00

We began by using the following model to analyze the data which includes covariate x_{2ij} in the model.

$$logit(p) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2ij} + \beta_1 z_{1ij} + \beta_2 z_{2ij} + b_i \tag{3.5}$$

3.2.1 GLMM in R

We fit the model in R using the package glmmML. This package uses maximum likelihood estimation via Laplace approximation to fit GLMs with random intercepts. This is very useful for clustered binomial data. Since our model is exactly that, a random intercept GLMM, this package is a perfect fit. R code for running this model is found in Appendix A.1.1. Table 3.5 shows results from this simulation example of a GLMM done in R.

Table 3.5: Results from the simulation example done in R providing maximum likelihood parameter estimates, estimated standard errors (SE), and a significance test of the hypothesis that the parameter equals zero for the model found in equation (3.5).

Parameter	True Value	Estimate	SE	Z	Pr($>\|z\|$)
α_0	0.5	0.463	1.019	0.454	0.650
α_1	2	4.080	1.592	2.563	0.010
α_2	0	1.279	1.278	1.001	0.317
β_1	1	5.010	2.039	2.457	0.014
β_2	-1	-4.335	1.886	-2.298	0.022

We then fit a second model, removing x_{2ij} from the model. Here we have the random intercept b_i, one level one and two level two explanatory variables. The data are clustered and the random intercept model takes into account the clustering of subjects.

$$logit(p) = \alpha_0 + \alpha_1 x_{1i} + \beta_1 z_{1ij} + \beta_2 z_{2ij} + b_i \tag{3.6}$$

Table 3.6: Results from the simulation example done in R providing maximum likelihood parameter estimates, estimated standard errors (SE), and a significance test of the hypothesis that the parameter equals zero for the model found in equation (3.6).

Parameter	True Value	Estimate	Std. Error	z value	Pr(>|z|)
α_0	0.5	0.89	0.57	1.57	0.12
α_1	2	2.44	0.56	4.32	0.00
β_1	1	0.08	0.64	0.12	0.90
β_2	-1	-0.69	0.68	-1.01	0.31

We can see from Table 3.6 that most of our estimates are reasonably close to the true values. All parameters are a little biased and standard errors are large, likely due to the small sample size.

3.2.2 Model Selection in R

Table 3.7 provides model selection results for the simulated data done in R. Here we find that AIC is indeed selecting the true model.

Table 3.7: Model selection results for the two models fit in R with the simulated data.

Model	mAIC	Δ_i	w_i
Equation 3.5	91.15	-1.53	0.00
Equation 3.6	89.62	0.00	1.00

3.3 GLMM in WinBUGS

We used WinBUGS to obtain Bayesian parameter estimates for the same simulated data. We were able to set up clusters of subjects like we did in R by using a start stop function. The priors used for this simulation were vague. For the α's and the β's we used flat normal priors. For the τ variable, which represents the precision in the random effects, we used a flat gamma prior. We chose these priors so they would not have much influence on our estimation.

The burn in period was a thousand iterations, we obtained 11000 samples, and the time to complete computations was only two seconds. The posterior mean estimates are reasonably close to the true values (Table 3.8). As point estimates, the posterior means are a little biased as was the case with R. The WinBUGS code for running this model is found in Appendix A.2.1. Table 3.8 shows results from this simulation example for the model containing the x_{2ij} covariate and Table 3.9 contains the results for the reduced model.

Table 3.8: Results from the simulation example done in WinBUGS providing posterior mean and standard deviation (SD) as well as the 2.5%, 50%, and 97.5% quantiles of the posterior distribution for the model found in equation (3.5).

Node	Posterior Mean	SD	2.5%	Median	97.5%
α_0	0.333	0.877	-1.327	0.299	2.178
α_1	3.447	0.892	1.993	3.341	5.441
α_2	0.913	1.087	-1.226	0.904	3.146
β_1	4.119	1.094	2.304	3.989	6.571
β_2	-3.277	1.139	-5.667	-3.216	-1.200
σ	1.402	0.806	0.306	1.262	3.328

Table 3.9: Results from the simulation example done in WinBUGS providing posterior mean and standard deviation (SD) as well as the 2.5%, 50%, and 97.5% quantiles of the posterior distribution for the model found in equation (3.6).

Parameter	Posterior Mean	SD	2.5%	Median	97.5%		
α_0	0.83	0.62	-0.30	0.80	2.14		
α_1	2.19	0.58	1.23	2.13	3.55		
β_1	1.79	0.67	0.51	1.77	3.16		
β_2	-1.07	0.68	-2.42	-1.06	0.25		
τ	2.03	3.38	0.14	0.85	11.87	1001	10000

3.3.1 Model Selection in WinBUGS

Table 3.10 provides model selection results for the simulated data done in WinBUGS. Here we find that DIC does select the true model.

Table 3.10: Model selection results for the two models fit in R with the simulated data.

Model	DIC	ΔDIC_i
Equation 3.5	86.78	-0.84
Equation 3.6	85.94	0.00

3.4 Comparison of R and WinBUGS

Both the classical and Bayesian methods provide reasonable estimates of the true values and are within one standard error of each other. This suggests the priors are not influencing the Bayesian estimates to a large degree. Further the DIC and AIC selection criterion chose the true model in the Bayesian and classical cases respectively.

Chapter 4

Modeling sharing behaviour

In May of 2008, community researchers (Erin Gibson and Heidi Exner) at AIDS Vancouver Island teamed up with members of the University of Victoria (Laura Cowen, Eric Roth, Junling Ma, and Pauline van den Driessche) to study the harm reduction practices amongst the clients of AIDS Vancouver Island's Street Outreach Services (AVI-SOS). AIDS Vancouver Island has had a needle exchange program running through their Street Outreach Services (AVI-SOS) since 1990. It has since been terminated, in May of 2008, at the fixed site but continues to run through a mobile service.

By examining the harm-reduction practices we were able to obtain the percentage of clients sharing needles, pipes, and drug equipment. This is valuable information as research has shown that sharing needles, pipes, and drug equipment leads to an increased risk of transmission of HIV and hepatitis C (Malliori et al., 1998).

The data consisted of 105 interviews with people who use injection drugs that were enrolled in the needle exchange program in Victoria, BC. The inclusion criteria were that the subjects had to be a registered user of the needle exchange and that they must have injected drugs within the past month. A registered user of the needle exchange was someone who had previously returned or received needles and was set up with an ID code. Subjects were given a 20 dollar honorarium to participate in the study. The interview was face to face and took about 30 minutes to complete. During the interview clients were read a questionnaire and their responses were recorded. Interviews occurred at AIDS Vancouver Island (AVI) in Victoria, BC.

Subjects were asked a variety of questions ranging from simple demographics such as sex and age to questions about their drug use network. A person's drug use network, in this case, consisted of people with whom subjects had talked about

using, buying, selling, or sharing drugs with during the past month. They were asked to list up to five people and answer questions pertaining to their relationship with this person. Questions such as type of relationship, age, sex, and length of relationship were asked and they had to fill in some descriptive statistics about the network members. Drug use sharing questions were also asked, such as whether they shared pipes or needles with a particular network member (See Appendix A.3 for a full copy of the questionnaire).

Out of the 105 interviews there were 71 males, 32 females, and 1 transgender. White was the prominent ethnicity at roughly 79%. There were 73 subjects who reported having a sex partner in their risk network. As for the sharing statistics, 24% of people reported sharing needles, 65% reported sharing pipes, and 57% reported sharing some other sort of drug equipment.

Interviewee variables were: Ethnicity, Education, Housing Status, Injection Frequency, Number of Injections, Sharing Needles, Sharing Pipes, Sharing Equipment, Fronted Drugs, Sex Partner, Pooled Money, and Gender (See Appendix A.3). Ethnicity was defined by five groups: white, black, hispanic, asian, and aboriginal. Education was broken into two sections of whether subjects went to and finished a certain level of education. Housing Status was defined as either homeless, own/rent, or subsidized. Injection Frequency came from how often in the past 30 days they had injected a particular drug. Number of Injections is how many injections were done in one day. Sharing Needles is if the client had shared needles with a member of his or her network. Sharing Pipes is if the client had shared pipes with a member of his or her network. Fronted Drugs is if the client had been given any drugs with the promise to pay back a person in his or her network. Pooled Money is if the client had pooled their money with a member of their network to buy drugs together. Sharing Equipment means the sharing of cookers, ties, water, cotton, or straws. Sexual Partner was if the subject had sex with a member of his or her risk network.

Table 4.1: Definitions of four variables used in the logistic regression analysis.

Variable	Definition
Sex Partner	0 = Did not have sex with someone in network
	1 = Had sex with someone in network
Share Equipment	0 = Did not share equipment with someone in network
	1 = Shared equipment with someone in network
Pool Money	0 = Did not pool money with someone in network
	1 = Pooled money with someone in network
Front Drugs	0 = Did not front drugs with someone in network
	1 = Fronted drugs with someone in network

Some aspects of these data have already been studied. Briefly, Exner et al. (2009) used exploratory analysis to delineate which factors people who use injection drugs worry about on a daily basis. They found that PWUID worry about contracting HIV/AIDS as well as overall personal security, and injection drug use specific risks (including overdose and vein collapse).

High risk injection practices were studied using a logistic regression analysis by Exner et al. (2010). They found that being homeless and the amount of time enrolled in the needle exchange program resulted in an interaction effect. Incidently, younger, male, homeless clientele that were enrolled in the needle exchange program for shorter periods were more likely to have higher risk scores.

Lindquist (2009) studied the dynamics of syringe sharing among pairs of injection drug users in Victoria and found that women were more likely to share with a sex partner. Using ordinary differential equations, Lindquist (2009) modelled pair formation and separation of injection drug users.

4.1 Logistic models for binary data

The first set of models we fit to the data were logistic regression models. The response in the first example was share needles. This is defined as

$$y_i = \begin{cases} 1 & \text{if the } i^{th} \text{ person shared needles with a member of their drug use network} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

The four variables that we used as predictors were sex partner, share equipment, pool money, and front drugs as defined in Table 4.1. These variables were suggested to be used by collaborators as they are viewed to be associated with needle sharing.

Table 4.2: Summary of the total number of individuals for each of the four predictor variables broken down by their specific 0 or 1 category (see Table 4.1) used in the logistic regression models: sex partner (SP), share equipment (SE), pool money (PM), and front drugs (FD).

Category	SP	SE	PM	FD
0	38	27	14	29
1	36	60	74	59

Before setting up the logistic model, we checked the contingency tables to determine if we had sparse data. Contingency tables with cells of low frequency may cause estimation problems for both the parameters and standard errors when fitting logistic regression models. Tables 4.3 - 4.6 show the 2x2 contingency tables of the predictor variables versus the response.

Table 4.3: Contingency table for variables share needles and sex partner.

	Share Needles	
Sex Partner	0	1
0	34	4
1	18	17

Table 4.4: Contingency table for variables share needles and share equipment.

	Share Needles	
Share Equipment	0	1
0	25	2
1	37	23

Table 4.5: Contingency table for variables share needles and front drugs.

	Share Needles	
Front Drugs	0	1
0	25	3
1	37	22

Table 4.6: Contingency table for variables share needles and pooled money.

	Share Needles	
Pool Money	0	1
0	13	1
1	49	24

From Tables 4.3 -4.6 we observe that some of the cell sizes are small. This might cause estimation problems when using the glm function in R to fit a logit model.

4.2 Model Selection

In order to obtain a parsimonious model we started with the null model and proceeded to add variables in a forward type approach by comparing the model's AIC_c value and assessing model weights. The null model is represented in Table 4.7 by a 1. R was used for all the logistic model analysis.

Table 4.7: Results of fitting 5 logistic models including Akaike's Information Criterion, ΔAICc, and model weight (w_{ic}). Variables included in the models are sex partner (SP), shared equipment (SE), front drugs (FD), and pooled money (PM). The null model is represented with a 1.

Model	Variable	AICc	ΔAICc	w_{ic}
4	SP	78.18	0.00	1.00
2	SE	98.26	20.07	0.00
3	FD	101.12	22.94	0.00
5	PM	103.78	25.60	0.00
1	1	106.40	28.22	0.00

Table 4.7 shows that the model with the lowest AIC_c contains SexPartner as a covariate. Therefore, we added SexPartner to the model and continued adding variables if the AIC_c value was reduced. In 4.8 shows a list of models that were fit.

Table 4.8: Logistic regression models fit to the AVI data with variables sex partner (SP), share equipment (SE), pool money (PM), and front drugs (FD).

Model	Variables	AICc	ΔAICc	w_{ic}
7	SP + SE + PM + FD	67.14	0.00	0.28
8	SP + SE + PM + FD + SP x SE	69.01	1.87	0.11
10	SP + SE + PM + FD + SP x FD	69.17	2.03	0.10
12	SP + SE + PM + FD + SE x FD	69.26	2.12	0.10
13	SP + SE + PM + FD + PM x FD	69.39	2.25	0.09
9	SP + SE + PM + FD + SP x PM	69.39	2.25	0.09
11	SP + SE + PM + FD + SE x PM	69.39	2.25	0.09
5	SP + SE + PM	70.20	3.06	0.06
6	SP + SE + FD	71.03	3.89	0.04
2	SP + SE	73.23	6.09	0.01
4	SP + PM	74.03	6.90	0.01
3	SP + FD	75.32	8.18	0.00
1	SP	78.18	11.04	0.00

The final model with the lowest AICc contained the covariates sex partner, share equipment, front drugs and pool money (Model 7). Table 4.9 contains the parameter estimates of this model using R software.

Table 4.9: Parameter estimates and standard errors of the logistic model (Model 7) containing covariates sex partner (SP), share equipment (SE), front drugs (FD), and pool money (PM).

Variable	Estimate	Std. Error
(Intercept)	-22.96	1545.27
SP	2.18	0.72
SE	2.30	1.16
PM	17.46	1545.27
FD	1.94	0.93

Looking at Table 4.9 one thing to note is that for the intercept and variable pool money the standard errors are extremely large. This is likely caused by the low cell value (of 1) shown in Table 4.6.

4.3 Modeling sharing behaviour with multi-level models.

Building on the logit models of the previous section we modeled the dyadic nature of the AVI data using multi-level models. The level one explanatory variables are descriptive about the client who we interviewed. Some of these examined were age, sex, and housing status. The level two explanatory variables come from the individuals representing the network. These consisted of sharing equipment, had sex with, and fronted you drugs. Definitions for these variables are found in Table 4.10. In Table 4.11, the distinct numbers in the ID column represent the interviewee. The number of times that distinct number is repeated represents the number of members in his or her network.

Table 4.10: Definitions of three level two variables used in the multi-level models.

Variable	Definition
Had sex with	0 = The network individual did not have sex with the interviewee
	1 = The network individual had sex with the interviewee
Share Equipment	0 = The network individual did not share equipment with the interviewee
	1 = The network individual shared equipment with the interviewee
Front Drugs	0 = The network individual did not front drugs to the interviewee
	1 = The network individual fronted drugs to interviewee

Table 4.11: A sample of the AVI data containing variables ego (ID), sex of the ego (Sex), age of the ego (Age), housing status of the ego (HS), had sex with a member of his or her network (HSW), fronted drugs to or from a member of his or her network (FD), and shared equipment with a member of his or her network (SE).

ID	Sex	Age	HS	HSW	FD	SE
1	0	34	1	0	1	1
1	0	34	1	0	1	1
2	1	19	1	1	0	1
3	1	29	2	0	0	1
4	0	31	0	1	1	0
4	0	31	0	1	1	0
4	0	31	0	1	1	0
4	0	31	0	1	1	0
4	0	31	0	1	1	0

The AVI data are rich in explanatory variables but poor in amount of data. Because of this, we fit models with only a few explanatory variables; thus, we started by using the variables found to be significant in the logit models of the previous section. The following describes models in our candidate set.

Model 1:

$$logit(p) = \alpha_0 + \alpha_1 x_1 + \beta_1 z_1 + \beta_2 z_2 + b \tag{4.2}$$

where x_1 represents age of clients interviewed, z_1 represents had sex with, and z_2 represents shared equipment.

Model 2:

$$logit(p) = \alpha_0 + \alpha_1 x_1 + \beta_1 z_1 + \beta_2 z_3 + b \tag{4.3}$$

where x_1 represents age of clients interviewed, z_1 represents had sex with, and z_3 is fronted money to buy drugs.

Model 3:

$$logit(p) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 z_1 + \beta_2 z_2 + b \tag{4.4}$$

where x_1 represents age of clients interviewed, x_2 is gender of the subject, z_1 represents had sex with, and z_2 represents shared equipment.

Model 4:

$$logit(p) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 z_1 + b \tag{4.5}$$

where x_1 represents age of clients interviewed, x_2 is gender of the subject, and z_1 represents had sex with.

Model 5:

$$logit(p) = \alpha_0 + \alpha_1 x_2 + \beta_1 z_1 + \beta_2 z_2 + b \tag{4.6}$$

where x_2 is gender of the subject, z_1 represents had sex with, and z_2 represents shared equipment.

4.3.1 Results

All five models were compared using AIC calculated from the glmmML package used in R. This AIC value is sometimes called "Marginal AIC" and is calculated using the usual mAIC formulation $AIC = -2\log\theta + 2k$ where k is the number of fixed parameters in the model and the likelihood is the marginal likelihood (Vaida and Blanchard, 2005). As we are interested in inference concerning the population parameters this model selection criterion will suffice. If we are interested in parameters specific to the clusters (the random effects) then a conditional AIC where the effective degrees of freedom is estimated would be used (Vaida and Blanchard, 2005).

4.12 shows the model selection results for the five candidate models. Model 5 has the lowest mAIC value and we chose this as our best model. We note that Model 1 is very close in mAIC value, and typically one would use model averaging to obtain final parameter estimates. It is not clear how model averaging is applied in random effects models and this is left for future work. Thus we include Model 1 in our discussion of parameter estimates.

Tables 4.13 and 4.14 provide parameter estimates for Models 1 and 5 respectively.

Table 4.12: Model selection results for the multi-level models; k is the number of fixed effects parameters.

Model	mAIC	Δ	k
1	157.9	0.5	3
2	164.0	6.6	3
3	159.3	1.9	4
4	164.4	7.0	3
5	157.4	0.0	3

Table 4.13: Parameter estimates and standard error estimates for Model 1 with covariates age, had sex with, and shared equipment.

Parameter	Estimate	Std. Error
α_0	0.47	1.45
α_1	-0.01	0.03
β_1	2.41	0.65
β_2	-1.36	0.52

Table 4.14: Parameter estimates and standard error estimates for Model 5 with covariates gender, had sex with, and shared equipment.

Parameter	Estimate	Std. Error
α_0	-0.31	0.41
α_1	0.38	0.45
β_1	2.42	0.65
β_2	-1.37	0.52

We note that the parameter estimates and standard error estimates for covariates had sex with (β_1) and shared equipment (β_2) from Models 1 and 5 are very similar.

4.3.2 Odds

Based on the results from Model 5, we can see that people are more likely to share needles with someone they have sex with, regardless of gender. The estimated odds of sharing needles for males versus females is $exp(0.38) = 1.46$. The estimated odds of sharing needles for those who also have sex with each other is $exp(2.42) = 11.25$ times that of not having sex for females and 12.06 for males. The estimated odds of sharing needles for those who also share equipment with each other is $exp(-1.37) = 0.25$ times that of not sharing equipment. This suggests that people aren't as likely to share needles with people they share equipment with. Perhaps people would be inclined to sharing a tie but wouldn't share a needle.

4.4 Modeling sharing behaviour using Bayesian hierarchical models

Similar to the previous section we fit Bayesian hierarchical models to the AVI data using the same five models. Here the priors on the α and β parameters were vague normal priors. The random effects b were fit with a random normal and for the precision we used a hyper parameter, τ. We fit τ with a flat gamma prior. We modeled the binary response as binomial. The code for these model is given in Appendix A.2.2. Like the simulation study in chapter three we had 1000 iterations of burn-in and did 10000 posterior samples of our MCMC chain.

All five models were compared using the DIC provided by WinBUGS. Model 5 was chosen as the best fit model, as it had the lowest DIC value 4.15. WinBUGS results for Model 5 are given in 4.16.

Table 4.15: Model selection using DIC.

Model	DIC	ΔDIC
1	155.15	1.53
2	154.30	0.68
3	154.32	0.70
4	156.47	2.85
5	153.62	0.00

Table 4.16: Results from WinBUGS of Model 5 including posterior mean and standard deviation and the 2.5%, 50%, and 97.5% quantiles of the posterior distribution.

Parameter	Posterior Mean	SD	2.5%	Median	97.5%
α_0	-0.43	0.49	-1.46	-0.40	0.46
α_1	0.38	0.58	-0.78	0.37	1.56
β_1	2.68	0.75	1.39	2.62	4.33
β_2	-1.42	0.62	-2.68	-1.40	-0.24
τ	1.27	2.17	0.13	0.56	7.82

4.5 Conclusions

Based on results from both WinBUGS and R we would choose Model 5 as the best model from our candidate set. Looking at the covariates of this model we can draw some conclusions as to why this was chosen. This model is backed to some degree by findings in the literature.

Males and females have different injection practices (Evans et al., 2003) in that females share both needles and drug preparation equipment more often than males. Evans et al. (2003) also found that females were more likely to share needles with their sex partner; further, females are injected by other individuals more often than males. Couples would be more inclined to share needles (Bryant et al., 2009), as was discussed previously thus we would expect the covariate 'have sex with' to influence sharing behaviours. If couples chose to share needles with their sexual partner, then sharing other injection equipment would not be as risky a behaviour. However, people are less inclined to share needles with someone in their network if they share equipment with them as well.

Chapter 5

Conclusions

The network data from AVI suggest, from both the Bayesian and classical maximum likelihood perspectives, that the important factors in sharing needles are: whether or not a person had sex with someone in their network, fronted drugs to or from someone in their network, and/or a person's gender. Logically these factors don't appear out of place. If a couple is sleeping with each other, they might think that because they know their partner they aren't worried about diseases. Sharing needles could be seen as an intimate gesture. Similarly, if someone is given drugs with the understanding that they pay later, they might be more inclined to share those drugs with whom they got them. A dealer could offer to split drugs with a person in order to recoup some of the loss. There could only be one needle and the person fronting the drugs is just focused on getting something. If someone pools money to buy some drugs they might be more inclined to share those drugs with the persons in the money pool.

5.1 Future work

We would like to expand the analysis done to study sharing pipes. Recent literature has shown that it is possible for the transmission of hepatitis C (HCV) from a host through crack paraphernalia (Fischer et al., 2008). Since the data are available this would provide good insight into other sharing practices. A comparison could be done to see if the same factors that are linked to sharing needles also apply to pipes.

5.2 Other Issues

We would have liked to have been able to fit larger more complicated models but the data were too sparse. When we tried to fit a more complicated model, R would produce error messages. In an ideal world we would have had a larger sample size to ameliorate the issue. This would have allowed us to fit models with interactions.

The data we collected would have been improved had every person answered all of the questions. When collecting the data, I felt that interviewees did not want to answer questions about all five people in their networks. They wanted to finish the survey as quickly as possible and the network section was at the end of the survey. Having an option to do 1-5 people made it easy for people to choose only one. Also, some people thought the network questions provided a way for researchers to identify the people they were talking about. They did not want to talk about people in their network for fear of breaking confidentiality, even though we did not collect personal identifiers. As for the validity of the data, I felt that people generally seemed like they were answering truthfully. Although, we have to assume truth for all questions answered, I noticed that for some questions people sometimes got annoyed and just provided an answer to move on.

Since the data suggests that people would share needles but not share equipment, it would be valuable to be able to take these results and ask the opinions of PWUID. The same was done in Exner et al. 2010 paper where PWUID linked injection practices to housing status.

Looking back at the history of the needle echange, it would have been useful to have done a follow up study with the same people. We could have seen if the patterns in sharing were changing with the fixed site being closed, moving to a strictly mobile needle exchange. However, having the needle exchange move to a mobile only operation it would have been difficult to locate the same people.

I think having a fixed site is a better way to enforce safe practices. Having a place where people can go and get cleaned up, something to eat, while getting safe injection equipment is vital.

Knowing from our best fit model that gender could affect sharing behaviour it might be a good idea to have separate ways of helping or reaching each gender. Health authorities could have different approaches or outreach practices to reach females compared to males. One such practice could be a time slot for females only to obtain clean needles. Alternatively, harm reduction education could be run differently for

females through shelters or community organizations where women feel welcome.

Appendix A

Questionnaire and Code

A.1 R Code

A.1.1 Simulation Data used in R

This code was used to create the simulated data set used in Chapter 3. Below we see the code used to generate the data, as well as the code to create the glmm models for R.

```
set.seed(1979)

n.subjects<-20
n.ans<-5
n<-n.subjects*n.ans
id <- factor(rep(1:n.subjects, rep(n.ans, n.subjects)))

x1.temp<-rnorm(n=n.subjects)
x1<-as.vector(t(outer(x1.temp,rep(1,n.ans))))
x2.temp<-rbinom(n=n.subjects,prob=0.5,size=1)
x2<-as.vector(t(outer(x2.temp,rep(1,n.ans))))

z1<-rbinom(n=n,prob=0.5,size=1)
z2<-rbinom(n=n,prob=0.5,size=1)
```

```
beta0<-0
beta1<-2
beta2<-2
alpha1<-3
alpha2<--2
sigma.sq.b<-2
b<-rnorm(n=n.subjects,mean=0,sd=sqrt(sigma.sq.b))

logit.p<-beta0+beta1*x1+beta2*x2+alpha1*z1+alpha2*z2 +
         as.vector(t(outer(b,rep(1,n.ans))))

p<-exp(logit.p)/(1+exp(logit.p))

Y<-rbinom(n=n,prob=p,size=1)

library(glmmML)
data.sim<-data.frame(id=id,shared=Y,x1 = x1, z1=z1, z2=z2)

fit1.logistic<-glm(shared~x1+x2+z1+z2,family=binomial,data=data.sim)

fit1.logistic.mixed<-glmmML(shared~x1+z1+z2,
family=binomial,data=data.sim,cluster=id)

summary(fit1.logistic.mixed)
summary(fit1.logistic)
#good table output for latex.
library(xtable)
xtable(data.sim)



#for winbugs
dput(data.sim$shared)
dput(data.sim$x1)
dput(data.sim$z1)
```

```
dput(data.sim$z2)

beg=seq(1,100,5)
end=seq(5,100,5)
dput(beg)
dput(end)
b=rep(0,20)
dput(b)
```

A.1.2 AVI data used in R

```
DATA2=read.csv("AVI data revised.csv", sep=",",header=T)
DATA2
library(glmmML)

attach(DATA2)
names(DATA2)
set.seed(79)
N=length(StudyIDCode)
n.subj=length(which(StudyIDCode!=1))
X1=AgeEgo
X2=GenderOfSubject
Z1=Hadsexwith
Z2=Sharedequipment
Z3=Frontyoudrugs

id.create=which(StudyIDCode!=1)
id=numeric(N)
for(i in 1:length(id.create)){
id[id.create][i]=i
}
for(i in 1:N){
if(id[i]==0){
id[i]=id[i-1]
```

```
}
}
beta0<-0
beta1<-.01
alpha1<-2
alpha2<--2
alpha3<-2
sigma.sq.b<-2
B<-rnorm(n=n.subj,mean=0,sd=sqrt(sigma.sq.b))

place=which(StudyIDCode!=1)
B1=numeric(N)
B1[place]=B
for (i in 1:N){
if(B1[i]==0){
B1[i]=B1[i-1]
}
}


logit.p<-beta0+beta1*X1+alpha1*Z1+alpha2*Z2+B1
p<-exp(logit.p)/(1+exp(logit.p))
Y<-rbinom(n=N,prob=p,size=1)

logit.p2<-beta0+beta1*X1+alpha1*Z1+alpha2*Z2+alpha3*Z3+B1
p2<-exp(logit.p2)/(1+exp(logit.p2))
Y2<-rbinom(n=N,prob=p2,size=1)
#########################################################################
data.sim<-data.frame(id=id,shared=Y, X1=X1, X2=X2, Z1=Z1, Z2=Z2, Z3=Z3)
fit1.logistic<-glm(shared~X1+Z1+Z2,family=binomial,data=data.sim)
summary(fit1.logistic)
library(glmmML)
fit1.logistic.mixed<-glmmML(shared~X1+Z1+Z2,
family=binomial,data=data.sim,cluster=id)
summary(fit1.logistic.mixed)
```

```
names(fit1.logistic.mixed)

fit2.logistic.mixed<-glmmML(shared~X1+Z1+Z3,
family=binomial,data=data.sim,cluster=id)
summary(fit2.logistic.mixed)

fit3.logistic.mixed<-glmmML(shared~X1+X2+Z1+Z2,
family=binomial,data=data.sim,cluster=id)
summary(fit3.logistic.mixed)

fit4.logistic.mixed<-glmmML(shared~X1+X2+Z1,
family=binomial,data=data.sim,cluster=id)
summary(fit4.logistic.mixed)

fit5.logistic.mixed<-glmmML(shared~X2+Z1+Z2,
family=binomial,data=data.sim,cluster=id)
summary(fit5.logistic.mixed)

fit6.logistic.mixed<-glmmML(shared~X2+Z1,
family=binomial,data=data.sim,cluster=id)
this=summary(fit6.logistic.mixed)

fit7.logistic.mixed<-glmmML(shared~X1+X2+Z1+Z2+Z3,
family=binomial,data=data.sim,cluster=id)
summary(fit7.logistic.mixed)

dput(data.sim)$shared)
dput(data.sim$X1)
dput(data.sim$X2)
dput(data.sim$Z1)
dput(data.sim$Z2)
dput(data.sim$Z3)
which(data.sim$X2=="NA")

library(xtable)
```

```
xtable(summary(fit1.logistic.mixed))

#########################################################################
data.sim2<-data.frame(id=id,shared=Y2, X1=X1, Z1=Z1, Z2=Z2, Z3=Z3)
fit2.logistic<-glm(shared~X1+Z1+Z2+Z3,family=binomial,data=data.sim2)
summary(fit2.logistic)
fit2.logistic.mixed<-glmmML(shared~X1+Z1+Z2+Z3,
family=binomial,data=data.sim2,cluster=id)
summary(fit2.logistic.mixed)


fit2.logistic.mixed$cluster.null.deviance-fit2.logistic.mixed$deviance
#[1] 0.4425934
fit2.logistic.mixed$cluster.null.df-fit2.logistic.mixed$df.residual
#[1] 1
dchisq(fit2.logistic.mixed$cluster.null.deviance-
fit2.logistic.mixed$deviance,fit2.logistic.mixed$cluster.null.df-
fit2.logistic.mixed$df.residual)
#[1] 0.4806176




End = cumsum(table(id))
Beg = 1+c(0,End[-length(End)])

Beg = as.vector(Beg)
End = as.vector(End)

dput(Beg)
dput(End)
```

A.2 WinBUGS Code

A.2.1 Simulation Model

Below we see the code used to create the glmm models for WinBUGS.

```
model
{
for( i in 1 : N ) {

for(j in Beg[i]:End[i]){
r[j] ~ dbin(p[j],1)
logit(p[j]) <- alpha0 + alpha1 * x1[j] +
beta1 * z1[j]   + beta2* z2[j] + b[i]
}



}
for( i in 1 : N ) {
b[i] ~ dnorm(0.0,tau)
}
alpha0 ~ dnorm(0.0,1.0E-1)
alpha1 ~ dnorm(0.0,1.0E-1)
beta1 ~ dnorm(0.0,1.0E-1)
beta2 ~ dnorm(0.0,1.0E-1)
tau ~ dgamma(1.0E-1, 1.0E-1)
sigma <- 1/sqrt(tau)
}



Data  list(
r = c(0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1,
1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
```

```
0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0),

x1= c(1.33400956244246, 1.33400956244246, 1.33400956244246, 1.33400956244246,
1.33400956244246, -0.6652044718667, -0.6652044718667, -0.6652044718667,
-0.6652044718667, -0.6652044718667, -0.971728453171647, -0.971728453171647,
-0.971728453171647, -0.971728453171647, -0.971728453171647, 0.281568461670842,
0.281568461670842, 0.281568461670842, 0.281568461670842, 0.281568461670842,
-0.0519030978663405, -0.0519030978663405, -0.0519030978663405,
-0.0519030978663405, -0.0519030978663405, -0.392768330035958,
-0.392768330035958, -0.392768330035958, -0.392768330035958, -0.392768330035958,
1.35625784148729, 1.35625784148729, 1.35625784148729, 1.35625784148729,
1.35625784148729, -2.26193117025432, -2.26193117025432, -2.26193117025432,
-2.26193117025432, -2.26193117025432, -1.48382240651445, -1.48382240651445,
-1.48382240651445, -1.48382240651445, -1.48382240651445, 0.183040484382072,
0.183040484382072, 0.183040484382072, 0.183040484382072, 0.183040484382072,
1.68624649467874, 1.68624649467874, 1.68624649467874, 1.68624649467874,
1.68624649467874, -1.10401268416574, -1.10401268416574, -1.10401268416574,
-1.10401268416574, -1.10401268416574, -0.868910560763743, -0.868910560763743,
-0.868910560763743, -0.868910560763743, -0.868910560763743, 0.65016923806339,
0.65016923806339, 0.65016923806339, 0.65016923806339, 0.65016923806339,
0.726253921175702, 0.726253921175702, 0.726253921175702, 0.726253921175702,
0.726253921175702, 1.02072068756142, 1.02072068756142, 1.02072068756142,
1.02072068756142, 1.02072068756142, -0.513005444599981, -0.513005444599981,
-0.513005444599981, -0.513005444599981, -0.513005444599981, -0.929700652719897,
-0.929700652719897, -0.929700652719897, -0.929700652719897, -0.929700652719897,
1.43476376309275, 1.43476376309275, 1.43476376309275, 1.43476376309275,
1.43476376309275, -1.53678250154046, -1.53678250154046, -1.53678250154046,
-1.53678250154046, -1.53678250154046),

z1 = c(1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0,
1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1,
0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0,
1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0),
z2 = c(1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0,
```

```
0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1,
0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1,
0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1,
0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0),
Beg=c(1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51, 56, 61, 66, 71, 76,
81, 86, 91, 96),
End=c(5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75,
80, 85, 90, 95, 100),
N = 20 )


Inits1  list(alpha0 = 0, alpha1 = 0, beta1 = 0, beta2 = 0, tau = 1,
b =c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
)
```

A.2.2 AVI Data in WinBUGS

```
model
{
for( i in 1 : N ) {

for(j in Beg[i]:End[i]){
r[j] ~ dbin(p[j],1)
logit(p[j]) <- alpha0 + alpha1 * x1[j] +
beta1 * z1[j]   + beta2* z2[j] + b[i]
}

}
for( i in 1 : N ) {
b[i] ~ dnorm(0.0,tau)
}
alpha0 ~ dnorm(0.0,1.0E-1)
alpha1 ~ dnorm(0.0,1.0E-1)
beta1 ~ dnorm(0.0,1.0E-1)
beta2 ~ dnorm(0.0,1.0E-1)
```

```
tau ~ dgamma(1.0E-1, 1.0E-1)
sigma <- 1/sqrt(tau)
z1[2]~dbin(0,1)
z1[20]~dbin(0,1)
z1[31]~dbin(0,1)
z1[34]~dbin(0,1)
z1[36]~dbin(0,1)
z1[38]~dbin(0,1)
z1[42]~dbin(0,1)
z1[43]~dbin(0,1)
z1[49]~dbin(0,1)
z1[53]~dbin(0,1)
z1[55]~dbin(0,1)
z1[64]~dbin(0,1)
z1[87]~dbin(0,1)
z1[93]~dbin(0,1)
z1[95]~dbin(0,1)
z1[104]~dbin(0,1)
z1[107]~dbin(0,1)
z1[108]~dbin(0,1)
z1[122]~dbin(0,1)
z1[127]~dbin(0,1)
z1[128]~dbin(0,1)
z1[134]~dbin(0,1)
z1[150]~dbin(0,1)
z2[3]~dbin(0,1)
z2[5]~dbin(0,1)
}



Data  list(
r = c(1, NA, NA, 0, NA, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1,
1, NA, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, NA, 1, 0, NA, 1, NA,
0, NA, 1, 1, 1, NA, NA, 0, 0, 1, 0, 0, NA, 0, 0, 0, NA,
```

```
0, NA, 0, 1, 1, 1, 0, 1, 1, 1, NA, 0, 1, 1, 0, 1, 1, 0, 0,
0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, NA, 0, 1, 1, 1, 0,
NA, 1, NA, 0, 0, 1, 0, 0, 1, 1, 1, NA, 0, 0, NA, NA, 1,
1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, NA, 0, 0, 0, 0, NA, NA,
0, 1, 1, 1, 1, NA, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, NA, 0),
x1= c(39, 39, 51, 51, 51, 60, 30, 38, 50, 50, 50, 50,
50, 57, 57, 50, 50, 47, 53, 49, 46, 46, 33, 47, 47,
47, 39, 39, 39, 50, 50, 39, 39, 39, 51, 45, 45, 53,
53, 53, 54, 45, 45, 41, 41, 41, 45, 49, 50, 42, 40,
40, 38, 38, 37, 40, 40, 50, 50, 44, 44, 59, 59, 59,
43, 49, 49, 49, 42, 42, 42, 44, 44, 37, 37, 37, 37,
37, 25, 34, 32, 36, 48, 39, 42, 42, 53, 43, 43, 43,
43, 43, 45, 45, 37, 37, 48, 48, 48, 35, 39, 42, 42,
57, 48, 48, 50, 45, 33, 34, 34, 37, 39, 46, 46, 60,
45, 45, 45, 53, 36, 35, 58, 58, 58, 45, 33, 33, 44,
44, 30, 30, 30, 50, 28, 28, 28, 28, 31, 40, 40, 40,
29, 46, 46, 46, 42, 42, 42, 35, 46),
z1 = c(0, NA, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
1, 1, 1, 0, NA, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, NA,
1, 0, NA, 0, NA, 0, NA, 0, 1, 1, NA, NA, 1, 0, 1, 0,
0, NA, 0, 0, 0, NA, 0, NA, 0, 1, 0, 1, 1, 0, 0, 0,
NA, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
0, 1, 1, 1, 0, 1, 1, NA, 0, 0, 0, 0, 0, NA, 1, NA,
0, 1, 0, 0, 0, 1, 0, 1, NA, 0, 0, NA, NA, 1, 1, 0,
0, 0, 1, 1, 1, 0, 1, 0, 0, 1, NA, 0, 0, 0, 0, NA,
NA, 0, 0, 0, 0, 1, NA, 0, 0, 0, 0, 0, 0, 0, 1, 1,
0, 0, 0, 0, 0, 0, NA, 1),
z2 = c(1, 1, NA, 0, NA, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0,
0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1,
1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0,
0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0,
1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1,
1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0,
```

```
0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1,
1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1,
1, 1, 1, 0, 1, 0, 1, 1),
Beg=c(1, 3, 6, 7, 8, 9, 14, 16, 18, 19, 20, 21, 23, 24, 27, 30, 32,
35, 36, 38, 39, 41, 42, 44, 47, 48, 49, 50, 51, 53, 55, 56, 58,
60, 62, 65, 66, 69, 72, 74, 79, 80, 81, 82, 83, 84, 85, 87, 88,
93, 95, 97, 100, 101, 102, 104, 105, 107, 108, 109, 110, 112,
113, 114, 116, 117, 120, 121, 122, 123, 126, 127, 129, 131, 134,
135, 139, 140, 143, 144, 147, 150, 151),
End=c(2, 5, 6, 7, 8, 13, 15, 17, 18, 19, 20, 22, 23,
26, 29, 31, 34, 35, 37, 38, 40, 41, 43, 46, 47, 48,
49, 50, 52, 54, 55, 57, 59, 61, 64, 65, 68, 71, 73,
78, 79, 80, 81, 82, 83, 84, 86, 87, 92, 94, 96, 99,
100, 101, 103, 104, 106, 107, 108, 109, 111, 112, 113,
115, 116, 119, 120, 121, 122, 125, 126, 128, 130, 133,
134, 138, 139, 142, 143, 146, 149, 150, 151),
N = 83 )


Inits1  list(alpha0 = 0, alpha1 = 0, beta1 = 0, beta2 = 0, tau = 1,
b =c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
)
```

A.3 AVI Questionnaire

PART I. GENERAL BACKGROUND DATA

The goal of this section is to determine who does, and doesn't use, the AVI-SOS on-site services.

Study Identification Code _____

1-1. Gender: Male Female Trans **Age in Years ____**

1-2. Ethnic Background: White Black Hispanic Asian Aboriginal

1-3. How many years have you lived in Victoria? _____

1-4. Where did you live before you came to Victoria? _____

1-5. Education: How long did you go to school?

Schools	**Go to**	**Finish**
Elementary School		
Middle School		
High School		
Community College		
Technical School		
University		
Apprenticeships		

Current Employment Status

1.6 Are you presently employed (have a straight job)? Yes No

1.7 If yes, what is the job ? _____

1.8 If no, how do you make money? _____

1.9 How much money do you make in a week? _____

1.10 What percentage of this money comes from a straight job? _____

1.11 How many places have you spent the night (slept) at in the past week?

Number ____

1.12 What is your present housing situation?

Own/Rent own home Yes No

Live in subsidized housing Yes No

Homeless Yes No

PART II - SUBSTANCE USE AND HISTORY

This section asks about drugs you may have used. For each drug I will ask you how old you were when you first tried it and how often you have used it in the last month, and the method you used (smoked, sniffed, injected).

SUBSTANCE	Age the first time you tried it	Never	Once a month	Twice a month	Once a week	Several times a week	Once a day	Several times a day	Method
		0	1	2	3	4	5	6	7
Tobacco									
Crack									
Cocaine									
Alcohol									
Marijuana									
Crystal meth									
Heroin									
Ecstasy/E									
GHB									
Ketamine/K									
LSD/Acid									
Peyote									
Methadone									
Magic Mushrooms									
Solvents									
Prescriptions/pills (specify)									
PCP									
Dipt									
DxM									

2-3. What is your preferred method of taking drugs (smoking, ingesting, sniffing, injecting)?

2-4. Has your preferred method changed over time? Yes __ No __

2-5. If yes, then please explain why.

PART 3 - PERSONAL USE BEHAVIORS

The following set of questions asks about your personal risk behavior patterns, limited to the last three months. Please answer yes or no, and tell us why in your personal case.

3-1. Do you :

Behavior	Yes	No	Why
Use Alone			
Do Testers			
Only buy from known dealers			
Swab before you shoot			
Filter your dope			
Use ties			
Reuse your own needle			
Help or give knowledge when you see others using unwisely			
Rotate veins			
Do dry shots			

Mix drugs			
Give out rigs			

3-3. Where do you inject : Public places ____ Private Places ____ Both ____

3-4. Do you have one or more places that you feel safe in when you use?
Yes ___ No ___

3-6. Have you ever injected someone else? Yes___ No____

3-7. If you don't inject yourself, who do you allow/ask to inject you?

3-8. If you don't inject yourself, have you ever tried to learn how?
Yes ___ No ___

3-9. How many times do you inject in a day?

3-10. How many times did you inject yesterday?

PART IV - PERSONAL RISK NETWORK

The next section seeks information on who you talk to about using, buying, selling, borrowing or selling drugs within the past month. To gather this information we ask you to construct your personal "Risk Network" by thinking of up to five people with whom you have shared this behavior in the past three months. We don't want to know anyone's real name.

4.1 NETWORK MEMBER	1	2	3	4	5
4.2 SEX					
4.3 AGE					
4.4 ETHNICITY (white, black, Asian, Aboriginal)					
4.5TIME KNOWN (in months or years)					
4.6 RELATIONSHIP (friend, lover, family member)					
4.7 USES NEP YES/NO IF YES, WHICH					
4.8 GAVE YOU NEEDLES					
4.9 SOLD YOU NEEDLES					
4.10 YOU GAVE HIM/HER NEEDLES					

QUESTION 4	CONTINUED				
4.1 NETWORK MEMBER	1	2	3	4	5
4.11 YOU SOLD HIM/HER NEEDLES					
4.12 SHARED NEEDLES					
4.13 SHARED COOKERS					
4.14 SHARED COTTON					
4.15 SHARED WATER					
4.16 SHARED STRAWS					
4.17SHARED TIES					
4.18SHARED PIPES					
4.19 FIXED FOR YOU					
4.20 DOCTORED THEM					
4.21 TRUST IN CASE OF OD					
4.22 FRONT YOU DRUGS					
4.23 HAD SEX WITH					
4.24USED CONDOMS Y/N					
4.25 TALKED ABOUT HIV					

4.26TALKED ABOUT HCV					

NOTES

PART V – RISK AND WORRY

In the following sections we ask you to assess different challenges you face everyday. Please use the numbering scheme to describe how much you worry about each item on a daily basis, and circle the number you think is most appropriate for you.

NUMBERING SCALE

1 = Never 2 = Once a month 3 = Weekly 4 = Daily 5 = All the time

RISK FACTORS

5.1. Overdose 1 2 3 4 5

5..2 Police Arrest 1 2 3 4 5

5.3 HIV Infection 1 2 3 4 5

5.4 Hepatitis C Infection 1 2 3 4 5

5.5 Sexually Transmitted Infections 1 2 3 4 5

5.6 Vein Damage 1 2 3 4 5

5.7 Being Robbed 1 2 3 4 5

5.8 Having a Place to Stay 1 2 3 4 5

5.9 Being Assaulted 1 2 3 4 5

5.10 Able to Buy Food 1 2 3 4 5

5.11 Are there any other risks you worry about? If so, please elaborate.

PART VI - SHARING SCENARIOS

In this section we want to present possible different situations that may arise and how the combination of factors that make up the situation influence the possibility of sharing syringes or injection equipment such as ties, cookers, water, etc.

To do this we ask about the following factors:

1) the *status* of the Needle Exchange Program, either "closed" or "open",

2) the *setting* for injecting – either a "public" setting like a park or abandoned building or a "private" setting like your, or a friend's home,

3) your injection *partner* – either a friend, or a sexual partner

4) HIV status, either your status, your partner's, or both

We first combine these factors in one single scenario, and then focus on each factor in that combination separately. For both approaches please use the following numbering scheme to say whether you consider your chances of sharing syringes and/or equipment (cookers, ties, cotton, etc.) as:
1 = Never 2 = Maybe 3 = Very Likely

Scenario 1. The NEP is closed, you are in a private setting, and your partner is a friend whose HIV status is unknown 1 2 3

Factors

NEP closed	**1**	**2**	**3**
Private Setting	**1**	**2**	**3**
Friend	**1**	**2**	**3**
HIV unknown	**1**	**2**	**3**

6.2 Scenario 2. You are in a public place, your partner is your lover who is HIV+, and the NEP is open 1 2 3

Factors

NEP open	**1**	**2**	**3**
Public Setting	**1**	**2**	**3**
Sexual Partner	**1**	**2**	**3**
HIV+ (lover)	**1**	**2**	**3**

6.3 Scenario 3. You are HIV+, your partner is a friend, the NEP is open and you are in a private setting. 1 2 3

Factors

NEP open 1 2 3
Private Setting 1 2 3
Friend 1 2 3
HIV+ (you) 1 2 3

6.4 Scenario 4. You are with your lover, the NEP is open, you are at home, and your HIV status is unknown. 1 2 3

Factors
NEP open 1 2 3
Private Setting 1 2 3
Sexual Partner 1 2 3
HIV unknown (you) 1 2 3

6.5. Scenario 5. You are with a friend, both of you don't know your HIV status, the NEP is closed, and you are at home. 1 2 3

Factors
NEP closed 1 2 3
HIV unknown (both) 1 2 3
Friend 1 2 3
Private Setting 1 2 3

6.6. Scenario 6. You are HIV+, you are with your lover in a public place and the NEP is closed. 1 2 3

Factors
NEP closed 1 2 3
HIV + (you) 1 2 3
Lover 1 2 3
Public 1 2 3

6.7 Scenario 7. Your friend is HIV+, you are HIV-, you are in a private setting and the NEP is closed. 1 2 3

Factors
NEP closed 1 2 3
HIV + (friend) 1 2 3
HIV – (you) 1 2 3
Friend 1 2 3
Private 1 2 3

6.8 Scenario 8. Neither you nor your lover knows your HIV status, you are in a private setting, and the NEP is open. 1 2 3

Factors

NEP open	**1**	**2**	**3**
HIV status unknown	**1**	**2**	**3**
Lover	**1**	**2**	**3**
Private	**1**	**2**	**3**

PART 7. CONCLUSIONS

7-1. To end this interview, is there anything you would like to add?

7-2. Are there any questions you'd like to ask us?

Date of Interview _____

Interviewer _____

Bibliography

[1] Agresti, A., Booth, J., Hobert, J. P., and Caffo, B. (2000) Random effects modeling of categorical response data. *Sociological Methodology*, **30**, 27–80.

[2] Amirkhanian, Y. A., Kelly, J. A., Kabakchieva, E., McAuliffe, T. L., and Vassileva, S. (2003) Evaluation of a social network HIV prevention intervention program for young men who have sex with men in Russia and Bulgaria. *AIDS Education and Prevention*, **15**, 205–220.

[3] Bailey, S. L., Ouellet, L. J., Mackesy-Amiti, M. E., Golub, E. T., Hagan, H., Hudson, S. M., Latka, M. H., Gao, W., and Garfein, R. S. (2007) Perceived risk, peer influences, and injection partner type predict receptive syringe sharing among young adult injection drug users in five U.S. cities. *Drug and Alcohol Dependence*, **91**, S18–S29.

[4] Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2008) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135.

[5] Bourgois, P., Prince, B., and Moss, A. (2004) The everyday violence of hepatitis C among young women who inject drugs in San Francisco. *Human Organization*, **63**, 253–264.

[6] Bruneau, J., Lamothe, F., Soto, J., Lachance, N., Vincelette, J., Vassal, A., and Franco, E. L. (2001) Sex-specific determinants of HIV infection among injection drug users in Montreal. *Canadian Medical Association Journal*, **164**, 767–773.

[7] Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theretic Approach*. Springer, New York, NY, second edn.

[8] Chitwood, D. D., McCoy, C. B., Inciardi, J. A., McBride, D. C., Comerford, M., Trapido, E., McCoy, H. V., Page, J. B., Griffin, J., and Fletcher, M. A. (1990) HIV seropositivity of needles from shooting galleries in south Florida. *American Journal of Public Health*, **80**, 1–3.

[9] Crowder, M. (1978) Beta-Binomial ANOVA for proportions. *Applied Statistics*, **27**, 34–37.

[10] Davey-Rothwell, M. A. and Latkin, C. A. (2007) HIV related communication and perceived norms: an analysis of the connection among injection drug users. *AIDS Education and Prevention*, **19**, 298–309.

[11] De, P., Jolly, A., Cox, J., and Boivin, J. (2006) Characterizing the drug-injecting networks of cocaine and heroin injectors in Montreal. *Canadian Journal of Public Health*, **97**, 207–209.

[12] Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., and Lee, R. S. (2009) Multilevel modeling: a review of methodological issues and applications. *Review of Educational Research*, **79**, 69–102.

[13] de Leeuw, J. and Meijer, E. (2008) *Introduction to Multilevel Analysis*. Springer, New York, NY.

[14] Dobson, A. J. (2002) *An Introduction to Generalized Linear Models*. CDC Press, Boca Raton, FL.

[15] Evans, J. L., Hahn, J. A., Page-Shafer, K., Lum, P. J., Stein, E. S., Davidson, P. J., and Moss, A. R. (2003) Gender differences in sexual and injection risk behavior among active young injection drug users in San Francisco (the UFO Study). *Journal of Urban Health*, **80**, 137–146.

[16] Exner, H., Gibson, E. K., Stone, R., Lindquist, J., Cowen, L., and Roth, E. A. (2009) Worry as a window into the lives of people who use injection drugs: a factor analysis approach. *Harm Reduction Journal*, **6**, 1–6.

[17] Exner, H., Gibson, E. K., Stone, R., Lindquist, J., Cowen, L., and Roth, E. A. (2010) A mixed methods approach to delineating and understanding injection practices among clientele of a Victoria, British Columbia needle exchange program. *Drug and Alcohol Review*, **30**, 360–365.

[18] Fischer, B., Powis, J., Cruz, M. F., Rudzinski, K., and Rehm, J. (2008) Hepatitis C virus transmission among oral crack users: viral detection on crack paraphernalia. *European Journal of Gastroenterology and Hepatology*, **20**, 29–32.

[19] Freeman, R. C., Rodriguez, G. M., and French, J. F. (1994) A comparison of male and female intravenous drug users' risk behaviors for HIV infection. *American Journal of Drug and Alcohol Abuse*, **20**, 129–157.

[20] Friedman, S. R. and Aral, S. (2001) Social networks, risk-potential networks, health, and disease. *Journal of Urban Health*, **78**, 411–418.

[21] Friedman, S. R., Curtis, R., Neaigus, A., Jose, B., and Des Jarlais, D. C. (1999) *Social Networks, Drug Injectors' Lives, and HIV/AIDS*. Springer, New York, NY.

[22] Goldstein, H. (2003) *Multilevel Statistical Models*. Wiley, Chichester, U.K.

[23] Gross, J. L. and Yellen, J. (2004) *Handbook of Graph Theory*. CRC Press, Boca Raton, FL.

[24] Heckathorn, D. D. (1997) Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, **44**, 174–199.

[25] Heckathorn, D. D., Semaan, S., Broadhead, R. S., and Hughes, J. J. (2002) Extensions of respondent-driven sampling: a new approach to the study of injection drug users aged 18-25. *AIDS and Behavior*, **6**, 55–67.

[26] Hurlbert, S. H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.

[27] Laireiter, A. and Baumann, U. (1992) Network structures and support functions: theoretical and empirical analyses. Veiel, H. O. and Baumann, U. (eds.), *The Meaning and Measurement of Social Support*, pp. 33–55, Hemisphere Publishing Corporation, New York.

[28] Lindquist, J. F. (2009) *Pair Formation and Disease Dynamics: Modeling HIV and HCV Among Injection Drug Users in Victoria, B.C.*. Master's thesis, Department of Mathematics and Statistics, University of Victoria, Victoria, B.C.

[29] Luke, D. A. (2004) *Multilevel Modeling*. Sage Publications, Thousand Oaks, CA.

[30] Luke, D. A. and Harris, J. K. (2007) Network analysis in public health: history, methods, and applications. *Annual Review of Public Health*, **28**, 69–93.

[31] Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000) Winbugs– a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

[32] Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. (2005) Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*, **19**, S67–S72.

[33] Malliori, M., Sypsa, V., Psichogiou, M., Touloumi, G., Skoutelis, A., Tassopoulos, N., Hatzakis, A., and Stefanis, C. (1998) A survey of bloodborne viruses and associated risk behaviours in Greek prisons. *Addiction*, **93**, 243–251.

[34] McCulloch, C. E. (2003) *Generalized Linear Mixed Models*. Institute of Mathematics and Statistics, Hayward, CA.

[35] McCulloch, C. E. and Searle, S. R. (2001) *Generalized, Linear, and Mixed Models*. Wiley, New York, NY.

[36] Miller, M. and Neaigus, A. (2001) Networks, resources and risk among women who use drugs. *Social Science and Medicine*, **52**, 967–978.

[37] Newman, M. E. J., Barabási, A., and Watts, D. J. (2006) *The Structure and Dynamics of Networks*. Princeton University Press, Princeton, NJ.

[38] Paccagnella, O. (2006) Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation Review*, **30**, 66–85.

[39] Scott, J. (2000) *Social Network Analysis: A Handbook*. Sage Publications, Thousand Oaks, CA.

[40] Snijders, T. A. B. and Berkhof, J. (2008) Diagnostic checks for multilevel models. *Handbook of Multilevel Analysis*, pp. 141–175, Springer, New York, NY.

[41] Stormer, A., et al. (2006) An analysis of respondent driven sampling with injection drug users (IDU) in Albania and the Russian Federation. *Journal of Urban Health*, **83**, 73–82.

[42] Unger, J. B., Kipke, M. D., Rosa, C. J. D., Hyde, J., Ritt-Olson, A., and Montgomery, S. (2006) Needle-sharing among young IV drug users and their social network members: the influence of the injection partner's characteristics on HIV risk behavior. *Addictive Behaviors*, **31**, 1607–1618.

[43] Vaida, F. and Blanchard, S. (2005) Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.

[44] Wedderburn, R. W. M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

[45] Whynot, E. M. (1998) Women who use injection drugs: the social context of risk. *Canadian Medical Association Journal*, **159**, 355–358.

[46] Wood, E., Tyndall, M. W., Spittal, P. M., Li, K., Hogg, R. S., Montaner, J. S. G., O'Shaughnessy, M. V., and Schechter, M. T. (2002) Factors associated with persistent high-risk syringe sharing in the presence of an established needle exchange programme. *AIDS*, **16**, 941–943.