A Study into Drug-trying Behaviour among Young People in England: Categorical Analysis Models in the Presence of Missing Data

by

Ho Hin Henry Chan

MSc Statistics, Lancaster University, 2012

BSc Mathematics, Hong Kong University of Science and Technology, 2011

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy from the Lancaster University

June 2018

A Study into Drug-trying Behaviour among Young People in England: Categorical Analysis Models in the Presence of Missing Data

by Ho Hin Henry Chan

MSc Statistics, Lancaster University, 2012

BSc Mathematics, Hong Kong University of Science and Technology, 2011

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy from the Lancaster University

June 2018

Abstract

This research reviewed the "Smoking, Drinking and Drug Use among Young People in England" 2010 survey (the Year 2010 Survey) study in terms of its data collection, processing and analysis. The research aim was to gain increased understanding of young people's drug-trying behaviour in England through appropriate handling of missing data, as well as, to build upon the previous work done, developing and applying statistical methodologies for analysis of multivariate categorical data collected by the Year 2010 Survey study.

The main work done in this research included: (1) modifying the original data set to arrive the useful working data set; (2) conducting exploratory data analysis with the working data set to identify direction for further empirical investigation; (3) properly handling the missing data problem in the working data set and (4) developing and applying advanced statistical methodologies to further analyse the working data set.

Apart from supporting the main findings of the Year 2010 Survey study that smoking, drinking and some drug-related socio-demographic covariates were

positively associated with the students' drug-trying behaviour, additional significant results found by the univariate logistic regression models, log-linear analysis models, two-parameter item response theory models and latent class analysis models reported that (1) the 15 drugs were highly and positively associated with each other and each drug exerted different extent of influences on the students' drug-trying behaviour and (2) generally, students' drug-trying behaviour could be further explained by numerous smoking, drinking and drug-related socio-demographic factors at different extent.

These additional findings contributed to a deeper understanding of the drug use problem, added evidence to the drug related research literature and provided helpful guidance on formulating policies to combat against drug use problem in England. Another contribution of this research was the development of a new methodology for backward elimination of latent class analysis models which provided a more thorough evaluation of the optimal number of latent class and covariate elimination from saturated model.

Acknowledgments

First and foremost, I would like to express my gratitude to my parents, Samuel and Anna, my sister, Lydia and my brother-in-law, Ning. They have given me their continuous support, encouragement, patience and understanding throughout the whole period of my doctoral degree study.

Second, I would like to sincerely thank my thesis supervisors, Dr. Gareth Ridall and Dr. Debbie Costain, who have made this dissertation possible. Without their wisdom, endless guidance and constructive advice, this dissertation would not be a reality.

Special thanks go to Prof. Kanchan Mukherjee for his support, encouragement and help. I am also deeply indebted to Dr. Louise Innes, Dr. Robert Blake and the internal and external examiners for their comments and suggestions made on the earlier drafts of this study.

Finally, a word of thanks to my colleagues and staff in the University of Lancaster who have generously accommodated me during my period of study at the University.

Declaration

I declare that this thesis results entirely from my own work and has not been submitted elsewhere for the award of a higher degree.

Signature:

Ho Hin Henry Chan

25 June 2018

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Drug use is a global problem and a long-standing issue for British society (Stimson (1987); McArdle (2004); Mold (2007); Niblett (2016)). Copps (2013) emphasized the seriousness of the problem in Britain by labelling it as "the addicted man of Europe"; outlining an increase in the number of people using various harmful drugs and yet who knew little of the damages that could be caused by those drugs. Over the years, drug use problem has impacted British society in various ways; for example: increasing the number of poisoning deaths, increasing the economic burden on drug addicts' families and society (Copello (2009); Copps (2013); Manders (2016)), and causing the health and social problems, such as disease transmission and growth in organised crime activities (Casey (2012); Copps (2013); Swiftl (2013)).

Regarding the policies on drug use, Copps (2013) further mentioned that the United Kingdom Government had endeavoured to combat against drug use and alcohol addiction problems, but barriers, such as established interests and funding cut, have impeded the government to effectively achieve its objective. Despite that, the United Kingdom Government has been implementing policies to combat against drug use problem (Stimson (1987); HM Government (2015); HM Government (2017)). To provide helpful guidance to the United Kingdom

Government on devising drug use policies, researchers usually rely upon surveys and statistical analysis, such as logistic regression models, with purposes to understand drug-trying behaviour, as well as to identify factors that are associated with drug use, for example, Fuller and Hawkins (2014). However, reviews of some drug-related studies indicated that a better approach, in terms of research methodologies to investigate drug-trying behaviour among young people, is needed and must be carried out, in order to enrich understanding of the drug use problem in England. With an objective to improve quality of future drug research studies, in this research, we focus upon development and application of advanced statistical methodologies to investigate drug-trying behaviour among young people in England. To achieve this research objective, a data set from one major survey series on drug use among young people, namely the "Smoking, Drinking and Drug Use among Young People in England" 2010 survey (the Year 2010 Survey) (Fuller et al., 2011), was utilised.

The "Smoking, Drinking and Drug Use among Young People in England" survey series, which was firstly carried out in year 1982 as a comprehensive biannual survey, is an annual survey that has been carried out jointly by the National Centre for Social Research and the National Foundation for Educational Research since 2000. The survey has been conducted to collect information about young people's behaviour and habits, in respect of smoking, drinking alcohol and drug use respectively. Please refer to Fuller et al. (2011) for the survey review.

The reported findings of the survey series have been considered by the United Kingdom Government when devising its policies on smoking, drinking alcohol and drug use among young people in the country (Department of Health, 2010). This annual survey has most recently been conducted by Statistics Team, NHS Digital (2017). For each annual survey conducted between the years 2010 and 2014 (Fuller et al. (2011); Fuller et al. (2012); Fuller et al. (2013); Fuller

and Hawkins (2014); Fuller et al. (2015)), non-responses (design-based and self-selected) and invalid or ambiguous responses existed in the data set. Potential reasons for the missing data may be the sensitive nature of the drug use questions posed, misunderstanding of questions and question ambiguity. Moreover, in the data analysis leading to the survey reports, potentially insufficient consideration has been given to handle the missing data issue. Specifically, simple methods, including treating missing categorical data as a separate category, were used to handle the missingness. In addition, further information regarding drug-trying behaviour among young people in England can be obtained through employing appropriate, possibly more advanced, statistical models in data analysis. For instance, one limitation of each annual survey is that for the drug-trying response variables, there was no consideration of the interactions of any sub-group behaviour. Another limitation is that the logistic regression models typically employed in data analysis investigated aggregation over all drug responses but did not consider each type of drug per se. As such, these inherent limitations may affect the robustness of the survey findings and may not have exploited sufficiently available information in the data collected.

Based upon the data collected by the Year 2010 Survey and built on its work done, the primary aim of this research is to gain increased understanding of drug-trying behaviour of young people in England by developing and applying advanced statistical methodologies to permit analysis of multivariate categorical data in the Year 2010 Survey study, in the presence of missing data.

1.1 A Primer on Drugs: Classification, Usage and Predictors

1.1.1 Definition and Classification of Drugs

According to the documents from the United Nations Office on Drugs and Crime (UNODC), in terms of international drug control, both the terms "drug" and "narcotic drug" are defined as "any of the substances listed in Schedule I and II of the 1961 Single Convention on Narcotic Drugs" (United Nations, 1961). These two schedules include, but are not limited to, cannabis, cocaine, heroin, methadone, morphine and opium. The term "narcotic drug" is used imprecisely to connote the term "illicit drug" in common parlance and legal usage (United Nations, 2016).

In Europe, drug classification varies among member countries of the European Union (European Monitoring Centre for Drugs and Drug Addiction, 2012a). For example, in the Netherlands, drugs are classified into soft and hard drugs (Government of the Netherlands, 2011), and in Ireland, drugs are classified into five schedules, where cannabis, LSD and MDMA are classified into the highest-level schedule (European Monitoring Centre for Drugs and Drug Addiction, 2012b), as being the most harmful drugs. The most prominent method of drug classification in the United Kingdom is using Schedule II of the Misuse of Drugs Act 1971, Chapter 38 (HM Government, 1971), in which drugs are classified into classes A to C, with class A represents the most dangerous drugs, and class C represents the least dangerous drugs. Class A drugs include crack, cocaine, ecstasy, heroin, LSD, magic mushrooms, amphetamines (if injected) and methadone. Class B drugs include amphetamines (if taken orally), cannabis and benzodiazepines (tranquillisers). Class C drugs include anabolic steroids and ketamine.

1.1.2 Drug Use Problem

Drug use problem has been widespread around the world, yielding a considerably large market value of illicit drugs (e.g. cannabis, cocaine, heroin, opium, methadone, morphine, amphetamine). In 2003, the global illicit drug retail market yielded a total of $ 321.6 billion US dollars. In the same year, the global illicit drug trafficking market reached a total of $ 94 billion US dollars, which was greater than the total amount of meat and cereal wholesale markets combined (United Nations Office on Drug Control and Crime, 2005). More recently, May (2017) estimated that the value of the global illicit drug trafficking market was between $ 426 billion and $ 652 billion in 2014, which suggested an increase in the value of the illicit drug market over the eleven-year period, from 2003 to 2014. Regarding the number and hence the proportion of people who used illicit drugs, the World Drug Report 2017 (United Nations Office on Drug Control and Crime, 2017) revealed an increase in illicit drug use from the year 2006 to 2015, with the figure rising from 208 million to 255 million over the period. In addition, there was an increase in the proportion of adults who used illicit drugs. It was estimated that in 2015, 5.3 % of people aged between 15 and 64 had used illicit drugs, compared to 4.9 % in 2006.

In addition to a general increase in the number of people using drugs, variation in the types of drugs used has been observed. For example, cannabis use became more prevalent in 2013 when compared to its use in 2009, which was reflected in its corresponding prevalence index that had increased from 100 in 2009 (2009 as the base year and 100 as the base index) to more than 105 in 2013 (United Nations Office on Drug Control and Crime, 2015). However, an opposite trend has been observed of using other drugs. For example, use of amphetamines and cocaine became less prevalent in 2013 when compared to 2009, with both indices dropped from 100 to below 95 over the period (United

Nations Office on Drug Control and Crime, 2015).

In the United Kingdom, drug use is also a long-standing problem. In 2013, new 'legal highs' were entering the drug market at the rate of one drug per week (Copps, 2013). On the contrary, according to the National Health Service, approximately 3.3 million adults in England aged 16-59 were using drugs in 2005, which dropped to around 2.9 million in 2011. Despite the decrease in the number of drug users from 2005 to 2011, the drug use among adults in England remained substantial (NTA, 2012). Drug use trends vary among different age groups. More recently, according to the Home Office, the percentage of adult users, aged 16-24, of drugs in England and Wales decreased gradually from around 30 % in 1996 to between 15 % and 20 % between the years 2012 and 2015, whereas the percentage of adult users, aged 30-59, of drugs remained similar (Lader, 2015).

Trends in the usage of various types of drugs among young people in England also vary. The percentage of the entire population in England that used cannabis dropped from 11 % in 2001 to 7 % in 2010 (NTA, 2012). Also, according to the National Health Service, there were 332,000 heroin and crack users and 130,000 people who injected drugs into their bodies in England in the year 2005/06. These figures dropped to 306,000 and 103,000 respectively in 2009/10 (NTA, 2012). In contrast, there was an increase in the proportion of 16-59 years-old adults that used Class A Drugs, from 2.7 % in 1996 to 3.2 % in the 2014/15 period (Lader, 2015). Also, an increase in the proportion of such adults using powder cocaine was found rising from less than 1 % in 1996 to 2.4 % in 2014/15 (Lader, 2015). According to the Home Office, there was a rise in the proportion of the population aged 16-59 who used anabolic steroids, from the year period 2004/05 to 2014/15 (Lader, 2015). Furthermore, the recent increase in the use of new psychoactive substances has also become a worrying phenomenon, despite

there was a slight decrease in the number of heroin or crack users in the past few years prior to 2013 (Copps, 2013).

Focusing upon young people, there has been a sustained prevalence of lifetime drug use among young people in the United Kingdom. Hibell (2011) pointed out that the lifetime use of illicit drugs (which included cannabis, amphetamines, cocaine, crack, ecstasy, LSD or other hallucinogens, heroin and GHB) in the United Kingdom was higher than the average of European Union (EU) (27 % for the UK versus 18 % for EU on average). In terms of lifetime use of cannabis, marijuana and hashish, figures representing the United Kingdom were higher than the EU average (UK: cannabis - 25 %, marijuana and hashish - 25 %, EU on average: cannabis - 17 %, marijuana and hashish - 17 %). In addition, the United Kingdom yielded an above-average percentage of students who had specifically used inhalants (10 % for the UK versus 9 % for EU on average), but it yielded a below-average percentage of students who had used tranquillisers or sedatives (3 % for the UK versus 6 % for EU on average). Besides, the report "Substance Misuse Among Young People 2011-12" reflected an increasing trend in the number of young people aged under 18 who sought specialist services due to cannabis addiction problems, between 2005/06 and 2011/12, from 9,000 to 13,000 (NHS, 2012).

There are different trends in the usage of various types of drug among young people. Recently, researches showed that cannabis is the most used drug among young people. From the "Smoking, Drinking and Drug Use among Young People in England 2013" report, among the 5,168 students aged 11-15, 11.3 % of them had used at least one drug during the year 2013, with 7.0 % of those students trying cannabis (Fuller and Hawkins, 2014). Moreover, according to Lader (2015), cannabis was the most commonly used drug among respondents aged 16-24 in England and Wales in year period 2014/15, with 16.5 % of the

respondents using it. Powder cocaine was the second most commonly used drug among those respondents in 2014/15, with 2.4 % of them using it.

1.1.3 Impact of Drug Use Problem

Drug use problem among the public has contributed to several social problems in the United Kingdom. Firstly, drug use problem has led to unnecessary deaths in England and Wales. According to Manders (2016), the number of drug poisoning deaths increased from 2,597 in the year 2012 to 3,744 in the year 2016. The number of deaths due to heroin and/or morphine increased from 579 to 1,209 between 2012 and 2016, and the number of deaths due to cocaine increased from 139 to 371 between 2012 and 2016. Secondly, drug use problem among the general public in the United Kingdom has posed an economic burden to the UK public, costing British taxpayers 15 billion pounds in one year (Copps, 2013).

On a more personal level, drug use problem has added a financial burden to the drug addicts' families. For heroin users or crack users or both, the cost to their families was estimated to be £9,497 annually in 2008 prices (Copello, 2009). In addition, drug use problem has affected the family's health and resulted in loss of the addicts' employment opportunities. The total annual cost among all British families was estimated to be £1.8 billion (Copello, 2009). Also, the total resource cost of NHS and local authorities was £747 million (Copello, 2009), which is a huge economic burden to the United Kingdom government.

In summary, drug use problem has caused health and societal issues for young people. For example, an increasing number of new drugs in the UK market has caused some young people to lose their "bladders" (Copps, 2013). A research on cannabis seizures found that cannabis is harmful to brain development, especially to those of young people with mental health issues (Swiftl, 2013). Also,

drug takers are suffered from poorer overall health (Casey, 2012). Taking drugs affect young people's employability as well, because most employers are not keen to employ drug takers (Casey, 2012).

Throughout the past two decades, the United Kingdom Government has implemented measures and strategies to combat against drug use problem among people (including young people) in the United Kingdom. From the Government (1998) report, the United Kingdom Government has outlined a ten-year plan with the following four aims in the strategy framework: (1) young people, (2) communities, (3) treatment and (4) availability. The Government's strategies are: (1) to prevent young people from abusing drugs; (2) to protect communities in the United Kingdom from drug-related crimes and behaviour; (3) to assist people suffering from drug problems and (4) to reduce the supply of illegal drugs in the market in the United Kingdom. In the plan, the Central Government acts as the enabler and coordinator which coordinated with Government anti-drug bodies, such as UK Anti-Drugs Coordinator and Deputy, and organisations at national and local levels translate the Government's aims into practice, as well as local drug-action teams, private sectors and media that penetrate through communities, parents and young people to spread the Government's message and vision of drug abuse among people. These four aims have been carried on by the current HM Government (2015). In addition, the United Kingdom Government has allocated more resources to combat against drug use problem by driving and throughout the Internet. The United Kingdom Government has also helped shaping international anti-illicit drug policy and practice, as well as leading in global illegal drug combating actions such as launching new initiatives on new psychoactive substances and coordinating with other countries to establish and promote anti-illicit drug research and analysis network (HM Government, 2015).

1.1.4 Risk Factors Related to Drug Use

Both the prevalence and impact of drug use have motivated researchers to conduct many surveys and studies in order to gain more understanding about the drug use problem and provide useful guidance to the United Kingdom Government to devise drug policies and strategies. Various reports and studies have suggested several risk factors that are associated with drug use. Firstly, according to a survey of 2,318 teenagers aged from ten to twelve from Glasgow and Newcastle (Mckeganey, 2004), drug-trying behaviour amongst teenagers was found associating with several family and peer factors. Family drug use was linked to teenagers' drug use, as 15.8 % of the respondents had families that had used drugs in the past, compared to 1.9 % of the respondents whose families did not use any drug (Mckeganey, 2004). This finding suggests that if a teenager's family used drugs, it is more likely for that teenager to try drugs.

Secondly, 16.8 % of the respondents that received low parental supervision had used drugs, compared to 1.6 % of the respondents who received high parental supervision (Mckeganey, 2004). This finding indicates the positive effect of parental supervision on drug-trying behaviour.

Factors such as smoking and drinking alcohol have regularly been found associating with drug use (Mckeganey, 2004). 19.7 % of the respondents drinking alcohol for at least a month tried drugs in the past, compared to 3.2 % of the respondents who did not drink alcohol (Mckeganey, 2004). 44.7 % of the respondents who smoked at least once a week used drugs in the past, compared to 3.2 % of the respondents who did not smoke (Mckeganey, 2004).

Moreover, a Europe-wide study (Vuolo, 2009) revealed that adolescents who knew hard drug users were more likely to use a drug in the previous month

(odds ratio of 5.605, standard error = 0.213). Furthermore, it was also found that adolescents living with parents who used drugs were more likely to use drugs (Copps, 2013). In addition, factors of "substance abuse", "mental health problems" and "criminality" among parents of adolescents were also found to influence an adolescent to try a drug (Gauffin, 2013).

Drug use among adolescents has also been found relating to age, gender and school failure. Firstly, the association between drug exposure and age as well as gender has been supported by Mckeganey (2004), which stated that being male and increasing age resulted in increased exposure to drugs. Secondly, the relationship of gender on drug use (hazard ratio of drug taking for males compared to females is 2.39, 95% confidence interval: (2.34, 2.45)) has been supported by Gauffin (2013), which revealed that males were 2.39 times more likely than females to use at least one drug. Furthermore, it was found by Fuller and Hawkins (2014) that in year 2013, a higher percentage of males than females in England had ever used any drug (16.6 % compared to 15.7 %) and individual drugs such as cannabis (9.1 % compared to 7.5 %) and cocaine (1.1 % compared to 0.7 %). Besides, the relationship of age on drug use has been supported by Fuller and Hawkins (2014) report, where increasing age from 11 to 15 was linked to increasing percentage of trying any drug in 2013 (from 3 % to 23.7 %), as well as individual drugs such as LSD (from 0.1 % to 0.9 %) and glue/gas/aerosols/solvents (from 2.1 % to 4.4 %). Finally, from Gauffin (2013) report, it was found that an adolescent suffering from school failure was 4.22 times more likely to use a drug (hazard ratio = 4.22, 95% confidence interval = (4.13, 4.31)).

More generally, other factors that have been found relating to drug use include: (1) poverty and unemployment (Ghodse, 2012) and (2) other drugs (Hale and Viner, 2013).

Although findings of prior research studies have provided information on several risk factors that contributed to drug use and the United Kingdom Government never stops implementing drug policies to combat against drug use problem, the trend and continued prevalence of drug use indicate that the issue has not yet been resolved. To address the prevalence of drug use issue, which has significant adverse social, economic and financial impact, as well as to gain fuller understanding and better investigation of the issue, it is anticipated that research efforts should be devoted in at least two dimensions: (1) continuous conduction of drug related studies to explore more insightful information about drug abuse phenomenon and behaviour and (2) review of prior research studies about drug use to identify limitations and weaknesses, and to develop and apply statistical methodologies to improve the quality of future drug related studies. The latter dimension is the focus of this research.

A usual method of conducting drug related research is through surveys. In the next two sections, the general issues in respect of survey studies and specific issues in respect of "Smoking, Drinking and Drug Use among Young People in England" survey series will be discussed.

1.2 Surveys: Questionnaire Design and Limitations

1.2.1 Brief Introduction of Surveys

Fink (2002) stated that surveys collect information about a specific group of population, to "describe, compare, or explain their knowledge, attitudes and behaviour". Mathers et al. (2007) also stated that survey "is a traditional way of conducting research", which is "useful especially for non-experimental descriptive designs that seek to describe reality". Moreover, surveys are adopted by

researchers, for instance, in societal and scientific aspects (Mathers et al., 2007). De Leeuw et al. (2008) provided an alternative objective of conducting surveys, that is "to obtain insight into the behaviour of the whole group of respondents".

There are several classifications of surveys. Mathers et al. (2007) suggested that surveys can be classified into two types: (1) cross-sectional surveys and (2) longitudinal surveys. Surveys that are carried out at only one time point are known as cross-sectional surveys, whereas those that are carried out over a certain period (in units of months or years) are known as longitudinal surveys. Mathers et al. (2007) further classified the longitudinal surveys into cohort surveys and trend surveys, where cohort surveys follow the same group of individuals over a certain time period and trend surveys ask different individuals the same questions at each time point, over a specified time period.

When investigating drug use among young people, usually either a longitudinal survey or a cross-sectional survey is adopted by researchers, depending on the objectives of the research.

1.2.2 Methods of Conducting Surveys and Construction of Surveys

Mathers et al. (2007) provided a comprehensive list of methods of collecting survey data: (1) face-to-face interviews; (2) telephone interviews and (3) questionnaires. Face-to-face interviews are labour intensive, but they can be the best way of collecting high-quality data. Face-to-face interviews are preferable for sensitive, but non-personal, subject matter (drug taking questions are personal, thus not suitable for face-to-face interview, as well as lengthy interviews). They are also preferable when the researchers need to cope with respondents with disabilities. Also, telephone interviews can be an effective and economical way

of collecting quantitative data, given that the ownership rate of telephones is high in the survey area and the questionnaire is short. However, according to Mathers et al. (2007), whilst telephone interviews are conducted within a limited period, face-to-face interviews have benefits that respondents are generally "more likely to complete the survey", once "they are committed", when compared to the telephone survey. In general, questionnaires are cheaper and quicker than face-to-face interviews, and are therefore more ideal for large and widely dispersed population.

Apart from the postal method, questionnaires can also be delivered via email and the Internet (Dillman et al., 2014). Moreover, surveys with questionnaires can be conducted in a specified venue, such as classrooms in secondary schools (Fuller et al., 2011). Recently surveys tend to combine several survey methods in a single survey for a reason: to increase the response rate and enhance the collection of survey data (Fink (2002); Dillman et al. (2014)).

Regardless of which survey method researchers are using, in most circumstances, a portion of respondents do not provide answers to some or all questions in a questionnaire. Missing data, also known as missingness, therefore exist in such circumstances. In Section 1.2.3, we discuss more the missingness problem.

1.2.3 Missingness Problem

Fink (2002) and Kang (2013) suggested that missing data occur in almost all survey research, "even in a well-designed and controlled study". According to Fink (2002), there are several causes that affect the level of missing data, which are also known as non-responses, including: (1) the nature of the population units; (2) the mode of data collection and (3) the fieldwork procedures together

with social and cultural factors. Major factors that correlate consistently with item non-response include the respondent's age and education, where elderly and less educated respondents tend to lead to an increased amount of missing data (Tourangeau and Yan, 2007). Also, one main cause of missingness problem is sensitive questions (Tourangeau and Yan, 2007).

According to Tourangeau and Yan (2007), sensitive questions tend to produce higher non-response rates than those on non-sensitive topics. Tourangeau et al. (2000) listed three distinct characteristics of sensitive questions: (1) "intrusiveness to privacy"; (2) "threat of disclosure" and (3) "social desirability". Survey questions about drug use and sexual behaviour have met all the three criteria of sensitive questions, so they are prone to missingness (Tourangeau and Yan, 2007).

Other reasons for non-response include: (1) the inclusion of 'do-not-know' questions (Sudman and Bradburn, 1974); (2) a respondent faced with a large number of questions (Weiner and Dalessio, 2006) and (3) refusal to participate and inability of the data collector and respondent to communicate, due to for example, language barriers (Fink, 2002). The existence of missing data may cause various issues in data analysis.

Most data analysis procedures are designed for complete data sets (i.e. data sets without any missing data) instead of data sets with missingness (Schafer and Graham, 2002). When these inferential methods are applied on data sets with missing data of which they are not dealt with beforehand, this may lead to "misleading inferences" (Carpenter and Kenward, 2013). Moreover, if the missing data are not handled properly, for example, by listwise or pairwise deletion (Kelejian (1969); Schafer and Olsen (1998)), information loss as well as "less efficient" estimates and less powerful "statistical tests", may result (De Leeuw,

2001). Missing data can also render the analysis invalid due to biased results (Kang, 2013). Details about missing data will be discussed in Chapter 4.

After presenting the general issues in respect of surveys, we introduce the smoking, drinking and drug use surveys in England in Section 1.3.

1.3 The Smoking, Drinking and Drug Use Survey in England

1.3.1 Overview of "Smoking, Drinking and Drug Use among Young People in England" Series

Among various drug use related surveys, the survey series of "Smoking, Drinking and Drug Use among Young People in England" are exemplars in terms of the scale of the survey, quality of study and extent of influence. The survey series began in 1982 by measuring the prevalence of smoking and smoking behaviour among young people in England. From 1988 onwards, the survey included alcohol consumption among young people, and from 1998 onwards, the survey also included the prevalence of drug use among young people. The survey series were carried out from 1982 to 1998 on a biannual basis. The survey series have then been carried out annually since 2000, jointly by the National Centre for Social Research and the National Foundation for Educational Research (Fuller et al., 2011), except the year 2015 survey, which was skipped due to an external sponsorship funding issue.

There are two aims of conducting the survey series. One aim of conducting the survey series is to address the Government of United Kingdom's concern "on the use of tobacco, alcohol and drugs" among young people in England

(Fuller et al., 2011). Another aim is to guide the United Kingdom Government's development and implementation of policy on smoking, drinking and drug use among young people, since the government recognises smoking, drinking alcohol and abusing drug as three of the seven most common primary causes of preventable deaths in England (HM Government, 2010). Thus, the findings reported in the survey series have been seriously considered by the United Kingdom Government (Department of Health, 2010).

Starting from 1998, each year's survey included a set of core questions covering students' current and past activities of smoking, drinking and drug use, consumption of cigarettes and alcohol drinks in the previous week prior to the study, as well as their awareness of and the availability of several specific drugs (Fuller et al., 2011). With effect from 2000, additional detailed questions were included in the annual questionnaire, with the emphasis alternating between smoking and drinking in one year and drug use the next (Fuller et al., 2011).

1.3.2 Overview of "Smoking, Drinking and Drug Use among Young People in England" 2010 Survey

In this research, the research aim is to review a previous research study about drug use, to identify its limitations and weaknesses and built upon its work done, to develop and apply statistical methodologies to gain increased understanding of drug-trying behaviour of young people in England. To achieve the research aim, we have chosen to critically review an annual "Smoking, Drinking and Drug Use among Young People in England" 2010 Survey (hereafter referred as the Year 2010 Survey) in terms of its data collection, data processing and data analysis with purposes to improve the quality of the survey study.

Apart from the reasons that the Year 2010 Survey is a comprehensive and per-

tinent related drug use study in England with data available at the time when this research began as well as there are potential rooms for improvement in respect of its data analysis conducted, another important reason to choose the Year 2010 Survey is that as the focus of the Year 2010 Survey was smoking and drinking, additional detailed questions were included in the questionnaires concerned smoking and drinking as opposed to drug use. Thus, selecting the Year 2010 Survey for this research would provide an additional benefit of fuller understanding of drug-trying behaviour among young people from further investigation of the associations between drug-trying response variables and the smoking, drinking and drug-related socio-demographic covariates.

In total, 246 schools throughout England participated in the Year 2010 Survey, and a total of 7,296 students completed the survey questionnaires. After the survey, the data in the collected questionnaires were double-checked by an external keying agency, and a report of findings of the Year 2010 Survey was then published (Fuller et al., 2011).

1.3.3 Overview of Findings of the Year 2010 Survey Report

In this section, since this research focuses on drug use among young people, we discuss the key findings from the drug use section of the Year 2010 Survey report (Fuller et al., 2011) (Serial number: 6883). Furthermore, with the logistic regression models employed in the report, we discuss how the smoking, drinking and drug-related socio-demographic variables were found to relate to drug-trying behaviour among students in the Year 2010 Survey. It should be noted that because of the new sample design of the Year 2010 Survey, selection weights were applied to the survey data by the researchers in data analysis. Details of the new sample design of the Year 2010 Survey will be discussed in Section 2.1.

The Year 2010 Survey report revealed that the prevalence of reported drug use among the students aged between 11 and 15 has declined over the ten years period from 2001 to 2010 as supported by three findings: (1) the proportion of the students who reported having taken drugs ever dropping from 29% in 2001 to 18% over same period in 2010; (2) the proportion of those students that took drugs in the last year (i.e. 2009) prior to annual survey dropping from 20% in 2001 to 12% 2010 and (3) the proportion of the students who had taken drugs in the last month dropping from 12% in 2001 to 7% in 2010. However, the age of the students was found positively associated with drug use among the students with a higher proportion of older students (15 years old) than younger students (11 years old) who reported taking drugs in each of the three circumstances: (1) taken drugs at least once; (2) taken drugs in 2009 and (3) taken drugs in the last month. No such similar pattern was seen in respect of gender of the students except a slightly higher proportion of male students (7%) than female students (6%) reported that they have taken drugs in the last month. Regarding the frequency of taking drugs, 2% of the students took drugs once within 2009, the year prior to the Year 2010 survey, 3% of the students took drugs in two to five occasions, 1% of the students took drugs in six to ten occasions, and 2% of the students took drugs in more than ten occasions. There was a higher proportion of older students (5% of 15 years old) than younger students (1% of 11 and 12 years old) who reported taking drugs at least once a month.

In terms of use of drugs, cannabis was the most widely used drug, with 8.2% of the students reported trying it in 2010. Among those students who had taken drugs in 2010, 71% of them had only taken one type of drug, 29% had taken two or more. The proportion of the students who had taken specific drugs in 2010 was observed to increase with age of the students. A higher proportion of older students (33% of 15 years old) than younger students (14% of 11 and 12 years old) was found to have taken two or more different types of drugs.

Apart from age, factors that were found to contribute to general drug use among young people were truancy and whether being excluded from school. The students who had truanted or been excluded from schools were more likely to take drugs more frequently than those who had not truanted nor been excluded from schools. It was found that 8% of the students who had truanted or been excluded from schools reported usually taking drugs at least once a month, compared to 1% of those who had not been excluded or truanted from schools.

Though overall there was a decline in the proportion of the students (28%) who reported having been offered drug in the survey, the proportion of the students who had been offered drugs increased with age that 49% older students (15 years old) reported they had been offered at least one drug when compared with 9% younger students (11 years old). Regarding sources of helpful information about drugs, the most likely sources of obtaining helpful information by the students were teachers (67 %), television (64 %) and parents (62 %). There were differences by age and by gender in respect of the reported sources of helpful information about drugs.

The main statistical method referenced in the "Drug Use" section of the Year 2010 Survey report was the logistic regression analysis (Fuller et al., 2011). The research team fitted a logistic regression model with a binary drug response, y_i, which recorded whether the student i had tried any drug in the year prior to the survey (i.e. 2009): $y_i = 1$ if the student i had tried any drug and $y_i = 0$ otherwise; $i = 1, \ldots, n$.

Model outputs were reported in the form of odds ratios relative to baselines of the corresponding factors. Odds ratios greater than 1 indicated increasing odds of a student trying drugs, whereas odds ratios less than 1 indicated re-

ducing odds of a student trying drugs. In the logistic regression model, only significant variables at 5% significance level were reported. The researchers used t-tests to determine the significance of covariate at each factor level, and reported 95 % confidence interval for the odds ratio of each factor level. If the 95% confidence interval did not include 1, the corresponding factor level of a covariate was significantly different from the reference category. This implied that the covariate was significantly associated with drug use in 2010 at 5 % significance level and vice versa. Covariates that were non-significant at all factor levels were not reported in the result of logistic regression.

When handling the missing values for each variable, the researchers did not exclude them but rather treated them as either a single category (missing category) for categorical variables or imputed the mean value of the respondents for continuous variables. The key covariates that were reported in the model were listed in Table 1.3.1.

Table 1.3.1: Table of Variables Adopted in the Logistic Regression Model of the Year 2010 Survey Report

Variable	*Type*	*Labels*
Response Variable		
Tried any drug in last year	Nominal	Yes, No
Student-level Variables		
Sex	Nominal	Boy (=0), Girl (=1)
Age	Linear	
Ethnicity	Nominal	White, Mixed, Asian, Black, Other
Smoking Status	Nominal	Non-smoker, Occasional-smoker, Regular-smoker
Whether Drunk Alcohol	Nominal	Never drunk alcohol, Drunk in previous week, Drunk, not in previous week
Ever Truanted	Nominal	Yes, No
Ever Been Excluded	Nominal	Yes, No
Receives Free School Meal	Nominal	Yes, No
Number of Books at Home	Nominal	
School-level Variables		
School Type	Nominal	Maintained schools, Academics, Independent
Sex of School Intake	Nominal	Mixed, Boys Only, Girls Only
Strategic Health Authority	Nominal	
% GCSE A*-C passes	Nominal	(in quantiles)
% students Eligible for Free School Meals	Linear	
% students with English as Additional Language	Linear	
Faith School	Nominal	None/Not known, Christian Denomination, Other Religion

According to Fuller et al. (2011), seven variables were found to be significantly related to drug use in 2010: (1) sex; (2) age; (3) ethnicity; (4) smoking; (5) drinking alcohol; (6) truancy and (7) exclusion. The odds ratios and the 95% confidence intervals for the significant variables were found to be as follows:

Firstly, girls were less likely than boys to have taken drugs in 2010 (odds ratio=0.74, 95% confidence interval = (0.58, 0.94)). Secondly, the odds of having taken drugs in 2010 increased linearly with age (odds ratio=1.13 for each addi-

tional year of age, 95% confidence interval = (1.02, 1.24)). Thirdly, students of Asian ethnicity were more likely than White students to have tried drugs in 2010 (odds ratio=2.10, 95% confidence interval = (1.34, 3.31)). However, when Mixed students, Black students and students from other ethnic backgrounds were compared to White students, no significant differences were observed. Moreover, the students who regularly smoke were more likely to have tried drugs when compared with non-smoking students (odds ratio=11.30, 95% confidence interval = (8.31, 15.35)). For occasional smokers, the odds ratio was 5.99 (95% confidence interval = (4.19, 8.56)). Also, the students who had drunk alcohol within a week before the survey were more likely to have tried drugs when compared with non-drinking students (odds ratio=6.94, 95% confidence interval = (4.97, 9.68)). Those who had drunk alcohol but not within a week before the survey were more likely to have tried drugs when compared with non-drinking students, but with a smaller magnitude of the increase in odds (odds ratio=3.32, 95% confidence interval = (2.48, 4.42)). The students who had ever played truant from school were more likely to have tried drugs than those who had not (odds ratio=2.44, 95% confidence interval = (1.81, 15.35)), and the students who had ever been excluded from school were more likely to have tried drugs than those who had never been excluded (odds ratio=1.70, 95% confidence interval = (1.26, 2.29)).

In summary, the key findings of the Year 2010 Survey revealed that: (1) the prevalence of drug taking behaviour among young people aged between 11 and 15 had declined from 2001 to 2010; (2) cannabis was the most widely used drug, and (3) the factors of sex, age, ethnicity, smoking, drinking alcohol, truancy and exclusion were associated with drug use among young people, albeit in different directions.

1.3.4 Limitations of the Analysis of the Year 2010 Survey

In the Year 2010 Survey, the researchers have methodically researched and planned their questionnaire design, data collection and analysis. However, there are several limitations relating to data analysis carried out for Fuller et al. (2011) report. One limitation of the logistic regression model in Fuller et al. (2011) study is that it only models the effect of covariates on a single response variable (i.e. whether the students had tried any drug or not) in one-way direction (i.e. how covariates affect response variables, instead of how response variables affect covariates). In other words, to investigate the two-way interactions between covariates and a response variable, two logistic regression models are required. Furthermore, it is considered that the data analysis can be further enhanced by employing more sophisticated statistical models to study the associations between drug-trying response variables and other related covariates, as well as the interactions among drug-trying response variables.

Another limitation is the insufficient consideration of the missing data, which are ubiquitous among survey data sets. On one hand, when publishing proportion tables and frequency tables for variable pairs, missing cases were ignored. On the other hand, in the logistic regression, missing data for each variable were treated directly in one of the following two methods: (1) as either a single category for categorical variables or (2) mean imputation, for continuous variables. In addition, the report did not explain in sufficient depth the reasoning of how those three types of missing values in the data set existed, as well as the consequent methods of treating these missing values other than ignoring them or setting them as mean values. As explained in the previous section, if the missing data in a data set are not adequately and properly managed, the robustness of the data analysis may be adversely affected. Statistical computational methods applied in the circumstances of ignoring missing data may lead

to misleading inferences (Sterne et al., 2009) and "biased estimates" (Kang, 2013).

In summary, in the Year 2010 Survey, though the data collection methodology of the researchers was robust and thorough, the data set is considered not exploited in sufficient depth. Also, treating the missing data as either a single category for categorical variables, or imputing the mean observed values for continuous missing values are considered not appropriate approaches to deal with the missingness problem and may induce bias in data analysis (Rubin, 2002).

1.4 Aim, Approaches and Expected Contributions of Research, Structure of Thesis

1.4.1 Aim of Research

The primary aim of this research is to gain increased understanding of drug-trying behaviour of young people in England, based upon the data collected by the Year 2010 Survey and built on its work done, by developing and applying advanced statistical methodologies to permit analysis of multivariate categorical data in the Year 2010 Survey, in the presence of missing data.

1.4.2 Approaches and Expected Contributions of Research

To achieve the aim of this research, the main approaches of the research are planned as follows:

(1) To tidy the original data set of the Year 2010 Survey by employing parsimonious number of variables into this research and combining excessive levels of some variables. The resultant data set (i.e. working data set) will be in a sim-

pler and more appropriate format, in order to investigate the interactions among the drug-trying response variables, as well as how the smoking, drinking and drug-related socio-demographic factors contribute to drug-trying behaviour, which has not been sufficiently investigated in Fuller et al. (2011) study.

(2) To deal with the missing data problem that existed in the Year 2010 Survey. Firstly, we will determine the type of missingness for each variable included in the working data set with explanations, and whether the missingness is ignorable. Secondly, we will apply various imputation methods to the working data set and compare the results from imputed data sets between imputation methods, to evaluate the difference in parameter estimates between imputation methods. For the 15 drug-trying response variables that will be described in Chapter 3, as well as other covariates (or explanatory variables), we will impute the missing groups by multiple imputation by chained equations (MICE). Alternatively, the drug-trying response variables will be imputed under fully Bayesian framework. As such, more unbiased values can be assigned to missing data based on other covariates.

(3) In our research, we will fit logistic regression models to explain the drug-trying response variables with individual drugs and other covariates. We will also run latent class analysis to model these drug-trying response variables and covariates. For purposes of selecting useful variables for the latent class analysis, we will employ the logistic regression models in our study, using Akaike Information Criterion (Sakamoto et al., 1986) for eliminating less essential related covariates relating to drug-trying response variables. To deal with the rare-case problem, we will investigate contingency tables between drug-trying response variables and smoking, drinking and drug-related socio-demographic covariates. If empty cells exist, then special methods to impute empty cells may be needed. As such, missing data can be assigned with more unbiased values

based on other covariates, which may contribute to more robust estimates and standard errors.

(4) Moving beyond the logistic regression models, we will apply various statistical models to estimate the associations between drug-trying response variables and other related factors, as well as the interactions among drug-trying response variables. In our analysis, we will employ log-linear analysis models, item response theory models, latent class analysis and K-means clustering to the working data set. The main purpose of employing various statistical models in this research is to analyse the drug-trying behaviour among young people in the Year 2010 Survey from different perspectives.

(5) To carry out variable selection, we will adopt backward elimination on statistical models employed for choosing the most parsimonious model. For purposes of combining results from imputed data sets, we will adopt Rubin's rule. For applying Bayesian approach to the analysis, we will determine the prior by sensitivity analysis under Bayesian framework, in order to determine the stability of estimate results for drug-trying response variables against choices of priors.

(6) Regarding latent variable models (i.e. item response theory and latent class analysis models), we will reduce the dimension of the drug-trying behaviour by employing a latent variable to represent the propensity for students to try drugs. Continuous latent variable model and discrete latent variable model are compared.

It is anticipated that this research will have the following three main contributions:

(1) Through employment of different imputation models, it will show proper

ways of dealing with missing data in survey research in general and in the Year 2010 Survey in particular.

(2) Through development and application of advanced statistical methodologies, it will show how to enhance the quality of data analysis in survey research in general and in the Year 2010 Survey in particular.

(3) The development and application of advanced statistical methodologies to the working data set of the Year 2010 Survey will provide a deeper understanding on the drug-trying behaviour of young people in England in terms of the interactions among drug-trying response variables and the associations between drug-trying response variables and the smoking, drinking and drug-related socio-demographic covariates.

1.4.3 Structure of Thesis

To deal with the missing data problem, as well as to identify factors that contribute to drug use among young people and to develop new methodologies to investigate associations between drug-trying response variables and covariates, this thesis is structured into two parts. After having introduced the drug abuse problem, the survey issues, the data source of this research as well as aim and objectives of this research study in Chapter 1, the first part of the thesis, Chapters 2 to 4, focuses on data cleaning, variable selection and imputation of missing data. In these chapters, we will focus on selecting variables that capture the most essential part of the questionnaire. We will focus on data processing and treating missing data through more sensible methods. Specifically, we will ask are there excessive levels in any variable. What are the sensible ways of categorising different types of missing data? Finally, we will focus on a robust method of imputing the missing data. The second part of the thesis, Chapters

5 to 7, focuses on the modelling of the imputed data set. In this part, we aim to ask how the drug-trying response variables are related to each other. Which smoking, drinking and drug-related socio-demographic covariates are associated with drug-trying behaviour? What statistical models are fitted on imputed data sets reflect about drug-trying behaviour of young people? What are the relationships between these statistical models?

1.5 Summary

This chapter has provided an overview of drug use problem, its adverse implication and previous research on risk factors that were associated with drug-trying behaviour among young people. Building upon such knowledge, an overview of the "Smoking, Drinking and Drug Use among Young People in England" 2010 survey study was carried out and its limitations were discussed. Finally, the aim and approaches of this research, expected contribution of this research, as well as the structure of this thesis were elaborated in this chapter. The aim of this research is 1. to review the "Smoking, Drinking and Drug Use among Young People in England" 2010 survey study (the Year 2010 Survey), in terms of its data collection, data processing and data analysis, 2. to identify its limitations and weaknesses, as well as 3. to build upon its work done to develop and apply statistical methodologies to permit analysis of multivariate categorical data in the Year 2010 Survey, in order to gain increased understanding of drug-trying behaviour of young people in England.

The next chapter provides a detailed discussion of the Year 2010 Survey, as well as data extraction, cleaning and variable selection, in respect of the data set of this research.

Chapter 2

Smoking, Drinking and Drug Use Survey 2010

As mentioned in Section 1.3.2, the Year 2010 Survey is the selected data source of this research. To understand more about the Year 2010 Survey, this chapter outlines the survey and questionnaire designs, and the data source of the Year 2010 Survey (Fuller et al., 2011). The methods adopted in pre-processing the data set of the Year 2010 Survey are also described in this chapter. The main purposes of pre-processing the Year 2010 Survey data set are to reduce the complexity of the original data set and to obtain a useful data set for this research (that is the working data set), of which the focus is on drug use among young people.

2.1 Survey Design

Four steps were included in the Year 2010 Survey (Fuller et al., 2011) to collect survey data: (1) selecting respondents; (2) issuing letters to respondents and arranging times to conduct the survey; (3) administering the questionnaires in classrooms and (4) performing validation tests on respondents.

The sample design of the Year 2010 Survey was firstly changed from a dis-

tribution across England proportionate to the distribution of survey population (adopted in previous years' survey) to a multi-stage sample design stratified by the 10 Strategic Health Authority (SHA) regions in England. The change in sampling methodology was intended to produce regionally representative samples in order to facilitate the production of regionally representative analysis while produce results comparable with previous years' survey ((Fuller et al., 2011)). The Year 2010 Survey commenced with two stages of student selection. In the first stage, 52 schools were chosen in each of 10 Strategic Health Authority (SHA) regions in England. A total of 520 schools in England were approached and invited, via letters and telephone calls, to participate in the survey. Four schools approached were later found to be not eligible due to an insufficient number of students and were, thereby, removed from the study. In the second stage, approximately 35 students were randomly selected from each of the remaining 516 schools, according to each school's self-sorted student register, with respect to tutor groups, classes or groups, within school years.

The selected students were provided with letters, issued from the National Centre for Social Research via their schools, asking for their parents' consent to participate in the survey. For every chosen school, a convenient time for the survey was negotiated among the interviewers of the National Centre of Social Research and the school committee.

To conduct the survey, according to Fuller et al. (2011), all the invited students who agreed to participate in the survey were "gathered together in a classroom", where they were monitored by an interviewer. Each student was given a questionnaire to complete within a period, called a fill-in period. During the fill-in period, participants were not allowed to chat among themselves nor looked at other students' answers. Moreover, they were informed by the interviewer and also through the questionnaire statements that their answers

would be completely confidential. To maximise the response rate, if four or more participants were absent during the first visit to a school, the interviewer visited that school for the second time. At this follow-up survey, participation progress was monitored and the same survey, for the previously absent participants, was conducted.

In the Year 2010 Survey, three factors were taken into consideration: (1) reliability of the participating students' answers; (2) honesty of the participating students and (3) accuracy of data collection through medical methods (Fuller et al., 2011). Firstly, to assess whether participants were honest in answering the questionnaire, researchers conducted saliva tests from students in half of the participating schools during the survey (Fuller et al., 2011). It was discovered that only several students yielded contradictory saliva levels against the smoking behaviour reported by themselves, indicating that most students were honest about reporting their smoking behaviour (Fuller et al., 2011). Secondly, the researchers inserted questions about a non-existent drug called semeron into- the questionnaire, in order to check if the students generally exaggerated their answers regarding drug use. It was found that only 13 out of 7,296 students reported that they had ever tried semeron. This indicated that most students did not exaggerate their drug use (Fuller et al., 2011).

In order to minimise recall bias of this investigation, the recall period of questions regarding the usual behaviour of the students related to alcohol and cigarettes were set to be within a week prior to the survey. One reason was that recalling the number of cigarettes smoked or the amount of alcohol drunk might be difficult for most students. Another reason was that the students' behaviour pattern might be discrete and "experimental". It could be that such behaviour pattern was caused not only by the students' own memories but also by their self judgement of their own memories. Also, the students' memories could not be

relied upon for a long period of time. Fuller et al. (2011) mentioned other sources of inaccuracy, including non-response bias and over and under-reporting, with the latter two factors potentially linked to the degree of social acceptance on smoking, drinking and drug use.

Furthermore, the new sample design of the 2010 Year Survey resulted in equal number of schools (52 schools) were selected in each of 10 SHA regions in England. Given the fact that the populations of the SHA regions varied, the probability that each student in the study would be selected was not the same across England. The survey data were therefore weighted (selection weights) by the researchers in order to correct the unequal selection probabilities among SHA regions (Fuller et al., 2011). Though it was understood that SHA regions, age and gender covariates were used to calculate selection weights, the calculation of the selection weights was not fully reported in the Year 2010 Survey Report nor could it be directly obtained from the researchers.

2.2 Questionnaire Design

The Year 2010 Survey questionnaire contained 238 questions spanning smoking, drinking, drug use and socio-demographics. Two types of questions were asked in the questionnaire: multiple choice questions and fill in the blank questions. The questionnaire began with six general questions, which captured the student's age, gender (Sex), school year (Syear), year and month of birth and ethnicity. The next 33 questions were about smoking habits, sources of cigarettes and the relationship between smoking and the respondents' peers. These questions were followed by 52 questions about drinking alcohol habits, sources of alcoholic drinks and the relation of alcohol drinking to people.

Among the 238 questions in the Year 2010 survey, 115 questions were specifically

related to drugs. These questions consisted of eight questions about each of the 15y separate drugs and other drugs as an independent category. These questions formed the largest section of the questionnaire. The 15 drugs in the survey were: cannabis, amphetamines, LSD, ecstasy, semeron, poppers, tranquillisers, heroin, magic mushrooms, methadone, crack, cocaine, ketamine, anabolic steroids and gas. These 115 questions were followed by four general questions about drugs. The following 25 questions were about socio-demographic factors, followed by two confirmatory questions about smoking frequency. The questionnaire concluded by asking the students if they had any other questions. The frequency table of each general classification of the questionnaire questions is shown in Table 2.2.1.

Table 2.2.1: Frequency Table of General Classification of Questionnaire Questions in the "Smoking, Drinking and Drug Use among Young People in England" 2010 Survey

Question Type	Frequency
General Question	6
Smoking-related	35
Drinking-related	52
Drug-related	119
Socio-demographic	25
Any other	1

Details about the classification of questions in the questionnaire are listed in Appendix A.1. The Year 2010 Survey adopted an internal routing system, in which respondents providing different answers in a question were directed to separate subsequent questions. For example:

Q9: Now read the following statements carefully and tick the box next to the one which best describes you.

Choice 1: I have never smoked → Q10

Choice 2: I have only ever tried smoking once → Q11

Choice 3: I used to smoke sometimes but I never smoke a cigarette now → Q11

Choice 4: I sometimes smoke cigarettes now but I don't smoke as many as one a week → Q11

Choice 5: I usually smoke between one and six cigarettes a week → Q14

Choice 6: I usually smoke more than six cigarettes a week → Q14

From the above example, if the students answered choice 1, they would be directed to question 10; if they answered choice 2, choice 3 or choice 4, they would be directed to question 11; and if they answered choice 5 or choice 6, they would be directed to question 14. As such, the students were directed to answer partial questions in the questionnaire that were applicable to them, skipping questions that were not. Finally, at the end of the questionnaire, a puzzle was provided to the students for entertainment after answering all the survey questions.

2.3 Open Data Source

A processed data set was uploaded onto the UK Data Service Website, formerly Economic Social Data Service. The data set, in SPSS format, is available on the website: `www.esds.ac.uk`. As mentioned in Section 2.1 that because of the new sample design of the Year 2010 Survey, selection weights were applied to the survey data by the researchers in data analysis. Nevertheless, in this study, we used mainly the unweighted data in data analysis rather than weighted data for the following main reasons:

(1) To achieve the aim of this study, we would develop and apply advanced statistical methodologies, such as log-linear analysis models (in Chapter 5), item

response theory models (in Chapter 6) and latent class analysis models (in Chapter 7) to further analyse multivariate categorical data collected in the Year 2010 Survey. Though Clogg and Eliason (1987) and Magidson (1987) had incorporated sampling weights into the maximum likelihood estimation for log-linear analysis, and Vermunt and Magidson (2005) suggested a method to incorporate sampling weights into latent class analysis, there were no methods to incorporate sampling weights into item response theory models. For maintaining consistency in data analysis under above-mentioned various advanced statistical methodologies as mentioned, we therefore did not incorporate selection weights (the calculation of them was not fully reported in the Year 2010 Survey Report) into our data analysis in this study.

(2) The 2010 Survey Report mentioned that the SHA regions were used in stratifying samples in order to facilitate the production of regionally representative analysis. Thus, selection weights were incorporated to correct the unequal selection probabilities among SHA regions. Nevertheless, the primary aim of this study is to gain increased understanding of drug-trying behaviour of young people in England rather than in each SHA regions. In such situation, according to Stapleton and Kang (2016), strategically, without access to multilevel software that can accommodate the sampling weights, we might consider including stratification variables (i.e. SHA regions, age and gender in this study) as independent variables in our data analysis. We therefore did not include selection weights in our data analysis but as a remedy, we included SHA regions, age and gender as independent covariates in our data analysis.

(3) Stapleton and Kang (2016) examined the design effects of five public-released data sets from the National Centre for Education Statistics (NCES) of ignoring the sampling design, and reported empirical findings that there were only minor effects of ignoring the sampling design and no differences in inferences would be

made. Also, in the Year 2010 Survey Report, some key survey estimates showed greater changes from 2009, while continuing established trends. Further analyses carried out by the researchers, including comparison of key estimates with and without selection weights, true standard errors and confidence intervals between Year 2009 Study and Year 2010 Study for these key estimates, did not indicate any reason to suggest that the changes in these key estimates were due to the change in sample design or the consequent selection weighting (Fuller et al., 2011). It was therefore believed that ignoring selection weights might not cause any significant statistical effect in our data analysis. Vermunt and Magidson (2007) also suggested that if the variables used to construct the sampling weighting do not affect the measurement part of the model, then we should use unweighted analysis rather than the weighted analysis.

The potential implication of using unweighted data in this study will be discussed in Section 8.8.1.

2.4 Data Processing

Examining the original data set of the Year 2010 Survey, a few issues associated with the data set were discovered. Firstly, the original data set contains 536 variables. We focused on the selection stage upon the variables directly recorded from the questionnaire rather than the derived variables, such as cigarette smoking status and non-cigarette-smoking status (three categories), because the original variables directly recorded information from the answers of the students' survey questionnaire. Since the focus of our analysis is investigating factors that contribute to drug-trying behaviour among young people in England, we selected questions and variables that were related to drug-trying, such as status of smoking and status of drinking (Mckeganey, 2004). We also opted to select a parsimonious number of variables, in order to apply the simplest statistical

models that have good explaining power. Therefore, the process of combining a few variables, to form a single variable with more information, was undertaken.

Secondly, as described in Fuller et al. (2011) report, the original data set contains three types of missing data. Including all three missing data categories in this research analysis would not gain extra benefits in research investigation, but would cause greater difficulty in analysing the data. As such, the missing data were recorded, trimming down the number of missing categories from three to one. Further details about recoding of missing data will be discussed in Section 2.4.2.1

Thirdly, the missing percentages of several original variables are too high. According to the data set, since the survey questionnaire adopted internal routing, most of the missing data of these original variables were linked to the leading questions, implying that most missing data was due to missingness by design. In this research, the missing data of the variables that were chosen were checked with the leading questions, in order to obtain certain corresponding answers to such missing data.

Finally, several original variables yield too many distinct levels, which lead to the following potential problems: (1) In a contingency table between one of such variables and a drug-trying response variable, empty cells might result and (2) unnecessary levels might result in longer analysis time when carrying out a logistic regression analysis. As such, the levels in these original variables were collapsed by combining levels with similar log odds ratios, whilst maintaining logical separations. An overview about collapsing these levels is provided in Section 2.4.3.

To summarise, in order to reduce the excessive complexity of the original data

set based on the above-mentioned four main reasons, the original data set is needed to be modified into a manageable and usable working data set for this research.

2.4.1 Modifications to Working Data Set

In this section, we describe the selection of the variables on four aspects: (1) drug-trying response variables; (2) smoking variables; (3) drinking variables and (4) drug-related socio-demographic variables.

2.4.1.1 Drug-trying Response Variables

15 drug-trying response variables, that identified the drugs which the students had ever tried, were selected (i.e. DgTdCan, DgTdHer, DgTdCok, DgTdMsh, DgTdCrk, DgTdMth, DgTdEcs, DgTdAmp, DgTdLSD, DgTdPop, DgTdKet, DgTdAna, DgTdGas, DgTdOth, DgTdTrn), and they were named as DgTdCan1 for cannabis, DgTdHer1 for heroin, DgTdCok1 for cocaine, DgTdMsh1 for magic mushrooms, DgTdCrk1 for crack, DgTdMth1 for methadone, DgTdEcs1 for ecstasy, DgTdAmp1 for amphetamines, DgTdLSD1 for LSD, DgTdPop1 for poppers, DgTdKet1 for ketamine, DgTdAna1 for anabolic steroids, DgTdGas1 for gas, DgTdOth1 for other drugs, DgTdTrn1 for tranquillisers respectively. The questions relating to whether the students had heard of the drug (i.e. DgHdCan, DgHdHer, DgHdCok, DgHdMsh, DgHdCrk, DgHdMth, DgHdEcs, DgHdAmp, DgHdLSD, DgHdPop, DgHdKet, DgHdAna, DgHdGas, DgHdOth, DgHdTrn) were not recorded, because variables capturing whether students had heard of a specific drug were deemed closely associated with the main response variables (whether they had tried that specific drug). If the students had not heard of a drug, then they were assumed to have never tried that drug. If the students were asked if they had ever heard of a drug, and they either answered 'Don't know' or refused to answer the question, then the response to the drug-trying

response variable was recorded as missing. If the students were asked if they had tried the same drug, and they either answered 'Don't know' or refused to answer the question, then the corresponding drug- trying variable was recorded as missing. The questions about the drug semeron were ignored, since there were too few cases of trying semeron (i.e. only thirteen cases) in the original data set to be used, and semeron is a fictional drug instead of an authentic one.

2.4.1.2 Questions relating to the addictive behaviour of smoking

In the Year 2010 survey, there were 103 variables recorded in respect of the addictive behaviour of smoking. The number of these variables was trimmed down to only 19 variables, which can be referred to Table A.2.1 in Appendix A.2 for the working data set. Reasons for trimming down the corresponding smoking variables are elaborated below.

For the three questions relating to family attitudes: (1) family's attitude to smoking (non-smokers) (CgFamN); (2) family's attitude to smoking (smokers) (CgFamS) and (3) family's attitude to smoking (secret smokers) (CgFamZ), a variable, CgFam1, was created to capture all the information about the family's attitude to smoking.

There were three questions relating to the severity of the smoking habit, including the cigarette smoking status (CgStat), the cigarette smoking status for irregular smokers only (CgIreg), and the total number of cigarettes smoked during the previous week in prior to the study, from Monday to Sunday (Cg7Mon, Cg7Tue, Cg7Wed, Cg7Thu, Cg7Fri, Cg7Sat, Cg7Sun). These questions were integrated into a single variable, CgStat1, which could be treated as an ordinal categorical variable. A variable (CgPe1) was used to capture the question of whether usually smoke packet cigarettes, roll-ups or both.

A question about ways of usually purchasing or obtaining cigarettes was adopted in data analysis. For that big question, there were 15 sub-questions asking the respondents how they obtained the cigarettes. These 15 sub-questions were categorized into three following groups: (1) group 1 - purchasing cigarettes through shops/machine/Internet (from supermarket, newsagent, garage, other type of shops, street market, machine, the Internet) (CgGetSup, CgGetNew, CgGetSho, CgGetMar, CgGetMac, CgGetInt); (2) group 2 - purchasing cigarettes through people (friends or relatives, or someone else) (CgGetFre, CgGetEls), and (3) group 3 - being given cigarettes by people or other sources (by friends, siblings, parents, someone else, or cigarettes in some other way) (CgGetGiv, CgGetSib, CgGetPar, CgGetElg, CgGetTak, CgGetOth). Each of these groups was treated as a separate variable, namely CgGet1 for Group 1, CgGet2 for Group 2 and CgGet3 for Group 3 respectively. The number of sources for each of these three variables was counted, and levels for each of these three variables based on the counts were classified, as well as alternative ways of obtaining cigarettes. These three variables could be treated as ordinal categorical variables. On the other hand, another variable, CgGet, was created, which determined whether the students obtained cigarettes through shops or people, or if they were given cigarettes by people. This created variable was derived from the three variables mentioned in this paragraph, and could only be treated as a nominal categorical variable.

There were eight sub-questions related to smokers that the students knew in a single big question (boyfriend or girlfriend, friends of same age, older friends, younger friends, parents or step-parent, sibling, other relatives, no friends or family) (CgPpGb, CgPpFrsa, CgPpFrol, CgPpFryo, CgPpPar, CgPpSib, CgPpOth, CgPpNo). The responses of these eight sub-questions were classified into three groups: (1) these smokers were other relatives; (2) these smokers were friends and (3) these smokers were family members. A derived variable of types

of people who know smoke cigarettes, namely CgPp1, was created of which the levels were determined by basing on which following group each student was classified: either other relatives, friends, family members or a mixture of these three groups.

For the two questions and the corresponding variables relating to smokers in house (whether people who lived with a student smoked inside the house) (CgWhoSmo, CgWhoHme), a combined variable, CgWho1, was created to capture both questions.

For the other two questions and the corresponding variables that were linked to the frequency of buying cigarettes from a shop, as well as how many peers of the students' age smoke, two separate variables, namely CgBuyF1 and CgEstim, were created to record them respectively.

There were several questions related to obtaining helpful information about smoking cigarettes from people (parents/ guardians, siblings, other relatives, friends, GP, teachers, other adults at school or police) (CgInPar, CgInSib, CgInRel, CgInFre, CgInGP, CgInTea, CgInAd, CgInPol) as well as several questions relating to obtaining helpful information about smoking cigarettes from the media (TV, radio, newspaper, the Internet, FRANK service, helpline) (CgInTV, CgInRad, CgInNews, CgInInt, CgInFRA, CgInHelp). A variable was created for the former set of questions in the same way as the variable related to people who the students knew smoke cigarettes, grouping these sub-questions into two groups: (1) obtaining information from parents and other relatives and (2) obtaining information from professionals and the police. This variable was named CgPe1. The same was done for the latter set of questions, grouping the sub-questions into two groups: (1) obtaining information through passive media and (2) obtaining information through interactive media. This variable

was named CgIn1.

Finally, two separate variables were created to capture issues about whether the students had lessons on smoking in the last twelve months (LsSmk), and whether the students currently smoked cigarettes (CgNow).

2.4.1.3 Questions relating to the addictive behaviour of drinking

In the Year 2010 Survey, there were 135 variables recorded in respect of the addictive behaviour of drinking. The number of these variables was trimmed down to only 21 variables, which can be referred to Table A.2.2 in Appendix A.2, for the working data set. The trimming down process is described below.

Firstly, a binary variable was created, which captured whether a student had ever drunk alcohol (AlEvr). Secondly, there were several questions relating to the severity of the drinking habit, including the frequency of drinking alcohol (AlFreq) and the number of days of drinking in the preceding week (Al7Day1), in the survey. These questions were combined into one created variable (AlFreq2) which could be treated as ordinal categorical variables. Thirdly, a binary variable was created, which captured whether a student had been in a pub, a bar or a club in the evening in the four weeks prior to the survey (AlBnPub). A variable (AlLast) was used to capture the question about when students last used alcohol.

A variable was created, which captured how many acquaintances of own age drink (AlEstim). This variable could be treated as an ordinal categorical variable. A binary variable, which captured whether the students had lessons on drinking in the last twelve months, was created as well (LsAlc).

There were three questions related to family attitudes on how parents feel about their children drinking alcohol: (1) how parents feel about their children drink-

ing alcohol that was applied to non-drinkers (AlPar); (2) how do parents feel about their children drinking alcohol that was applied to drinkers they knew (AlParSt) and (3) how do parents feel about their children drinking alcohol that was applied to drinkers they knew (AlParKnw) separately. To capture information from these three questions, a derived variable was created (AlPar1), which captured all the data about the family's attitude towards drinking alcohol. This variable could be treated as an ordinal categorical variable.

Questions about the number of places a student purchased alcohol, as well as the number of sources of obtaining alcohol, were adopted. There were eight sub-questions that asked the students from where and from whom they purchased alcohol (pub or bar, club or disco, off-license, shop or supermarket, friend or relative, off the street, garage forecourt or someone else) (AlBuyPub, AlBuyClu, AlBuyOff, AlBuyShp, AlBuyFre, AlBuyStr, AlBuyGar, AlBuyEls). These eight questions were categorized into two separate groups (AlBuy1 and AlBuy2). The number of sources for each of these two variables was counted, and classified levels for each of these two variables based on the counts, and alternative ways of purchasing alcohol. These variables could be treated as ordinal categorical variables. On the other hand, another variable (AlBuy) was created, which determined whether the students purchased alcohol from shops or acquired it from people. This variable could only be treated as a categorical variable.

In addition, two separate questions about types of peers that the students used alcohol with, and where the students used alcohol were adopted. For the question about types of people that the students used alcohol with, the seven sub-questions (girlfriend or boyfriend, same sex, opposite sex, both sexes, guardians, siblings or other relatives, or other people) (AlUsGB, AlUsFreS, AlUsFreO, AlUsFreB, AlUsPar, AlUsSib, AlUsOth) were classified into two groups: (1) other people and friends and (2) family members. A derived variable (AlUs1)

was created to capture types of people that students used alcohol with, and the levels were determined basing on which group each student was classified: either other people and friends, and family members. On the other hand, for the question about types of places the students used alcohol with, the seven sub-questions (pub or bar, club or disco, party, home, someone else home, street or somewhere else) (AlUsPub, AlUsClu, AlUsFre, AlUsHom, AlUsOHm, AlUsStr, AlUsEls) were classified into four groups: (1) pubs; (2) home/party; (3) stranger's place/public outdoor area and (4) a mixture. A variable (AlUs2) was created, which captured all the data about places a student usually used alcohol.

There were eight sub-questions in a single question asking about issues when drinking alcohol in the last four weeks. These eight sub-questions (had argument, had fight, felt ill or sick, vomited, taken to hospital, lost money or belongings, clothes, belongings damaged, or trouble with police) (Al4WArg, Al4WFig, Al4WIll, Al4WVom, Al4WHos, Al4WLst, Al4WDam, Al4WPol) were classified into two groups: (1) health issue and aggressive issue and (2) other issues. A variable (Al4W1) was created to indicate which group each respondent belonged to: (1) never drank; (2) drank but no issues; (3) health issues; (4) aggressive issues and other issues, and (5) both. There were also eight sub-questions within a big single question, asking about why the students thought about the reasons for the people of the same age to drink (relax, feel more confident, to be sociable with friends, bored, look cool, forget problems, for a rush or pressure from friends) (AlWhyRel, AlWhyCon, AlWhySoc, AlWhyBor, AlWhyCoo, AlWhyFgt, AlWhyRsh, AlWhyPre). These eight sub-questions were categorized into two groups: (1) to feel better and (2) to socialise. Another variable (AlWhy1) was created to indicate which group each student fell into in respect of the reason: (1) to feel better; (2) to socialise and (3) both. This variable could then be treated as a nominal categorical variable.

For the two questions and corresponding variables relating to drinking within their household (whether people who lived with a student drank inside the house) (AlWhoHme, AlWhoDr), a combined variable (AlWho1) was created to capture data of both questions.

There were eight questions relating to the source of obtaining helpful information about drinking alcohol from people (parents/guardian, siblings, other relatives, friends, GP, teachers, other adults at school or police) (AlInPar, AlInSiB, AlInRel, AlInFre, AlInGP, AlInTea, AlInAd, AlInPol), as well as six questions relating to getting helpful information about drinking alcohol from media (TV, radio, newspaper, the Internet, FRANK service, helpline) (AlInTV, AlInRad, AlInNews, AlInInt, AlInFRA, AlInHelp). A variable (AlPe1) was created for the former set of questions in the same way as we did for drinking alcohol, and the sub-questions were grouped into two groups: (1) parents and other relatives, and (2) professionals and police. By creating a variable (Alln1), a similar way was done for the latter set of questions, and the sub-questions were grouped into two groups: (1) passive media and (2) interactive media.

2.4.1.4 Drug-related Socio-demographic Questions

There were eight questions related to how the students gained knowledge about drug use from other people and from the media (parents/guardian, siblings, other relatives, friends, GP, teachers, other adults at school or police) (DgInPar, DgInSiB, DgInRel, DgInFre, DgInGP, DgInTea, DgInAd, DgInPol). A variable (DgPe1) was created for the first set of questions in the same method as the variable related to people who take drugs, and the sub-questions were grouped into two groups: (1) parents and other relatives and (2) professionals and police. Similar method applied to the media questions by grouping the six sub-questions (TV, radio, newspaper, the Internet, FRANK service, helpline) (DgInTV, DgInRad, DgInNews, DgInInt, DgInFRA, DgInHelp) into two groups: (1) passive

media and (2) interactive media by using a variable (DgIn1) to capture data of these sub-questions.

A variable was also created to capture how many acquaintances of the student's own age uses drug (DgEstim). This was an ordinal categorical variable. A binary variable that captured whether the students had drug education lessons in the last twelve months was also created (LsDrg).

Two variables were created, one was to capture how many books in the student's home (Books1) and another was to capture the age of the students (Age), which ranged from eleven to fifteen years old. These variables could be treated as ordinal categorical variables.

To capture the information about the gender of the students, a binary variable (Gender) was used. Another binary variable (FSM1) was adopted to record the question about whether the students had joined the free school meal (FSM) scheme was used to reflect the economic status of the student' families.

Moreover, in order to capture whether the students had ever played truant or had ever been excluded from school, two separate variables, namely Truant1 and ExclA1 were used respectively. The two variables could be treated as binary variables.

By the concept of extension of truancy and exclusion variables, two additional variables, namely TruantN and ExclAN1, were created to capture the students' frequency of playing truant and being excluded from school respectively. These two variables could be treated as ordinal categorical variables.

Finally, to incorporate Strategic Health Authority (SHA) regions in data analysis,

a variable, namely SHA, was created, which captured the ten SHA regions in England that were used to stratify the 7,296 students.

2.4.2 The Missing Data Problem

The original data set contained three types of missingness: (1) a question not answered or refused to answer by a student (coded as -9); (2) a question answered "don't know" or "can't tell" by a student (coded as -8) and (3) a question that was not applicable to a student (coded as -1). In this section, the general methods that are used to recode the missing data will be discussed, with the aim of downsizing the number of missing data categories to one.

2.4.2.1 Recoding of the Missing Data

For the missing data that were coded as (-9), they were all treated as missing, because no information could be obtained from this kind of missing data.

For the missing data that were coded as (-8), the corresponding question was examined to determine if this classification of missing data, "don't know" or "can't tell", could be regarded as a level in the subsequent variable. For example, in the case of creating a variable that described family attitudes toward smoking, the choices of the related variable were classified into three options: (1) against smoking; (2) for smoking and (3) neutral (between against option and for option). In this case, the students coded (-8) were treated to be in the middle (neutral) option, because they still answered as "don't know" or "can't tell", or simply ticked more than one box in any of these related questions about family attitudes toward smoking, and we were not sure if the families of those students clearly supported smoking or opposed to smoking. If the students did not tick proper boxes through the normal procedure, their answers were classified as "don't know". The same recoding strategy was applied to the vari-

able capturing how the students' parents/guardians feel about drinking alcohol.

For the missing data that were coded as (-8) in all other questions in which sufficient information could not be obtained to determine which valid option such missing response data could assume, the missing data that were coded as (-8) were treated to be missing.

For the missing data that were coded as (-1), which stood for "not applicable", the leading question was traced back to determine where those missing data should be recoded. For example, when the students were coded (-1) as responses for the variable "whether usually smoke packet cigarettes, roll-ups or both about equally", the leading questions, the question about cigarette smoking status and its subsequent question about cigarette smoking status for irregular smokers, were examined to determine how the "not applicable" responses for the question "whether usually smoke packet cigarettes, roll-ups or both about equally" were treated. When these students answered the question about cigarette smoking status or the subsequent question for irregular smokers, there were generally two scenarios listed as follows.

Scenario 1: Several students did not answer the question about cigarette smoking status or answered "don't know" for cigarette smoking status question;

Scenario 2: Several students answered "I have only ever tried smoking once" or "I used to smoke sometimes but I never smoke a cigarette now" for cigarette smoking status question. Other students answered "I have never tried smoking a cigarette, not even a puff or two" or "I did once have a puff or two of a cigarette, but I never smoke now" for cigarette smoking status question for irregular smokers. Also, some students did not answer the cigarette smoking status question for irregular smokers because they answered "I have never smoked" for cigarette

smoking status question.

After answering any of the two questions about smoking status, the students were then required to answer the following question - "whether usually smoke packet cigarettes, roll-ups or both about equally". There were three options provided in the questionnaire for the students to answer: (1) Cigarette from a packet; (2) Hand-rolled cigarettes and (3) both about equally.

Regarding the above question, for the students in scenario 1, their responses were treated as missing because there was no information or hint about which option these students should be allocated to. For the students in scenario 2, their responses were treated to be level 0: "never smoke now and usually", due to the questionnaire design that the students who matched the cases included in scenario 2 were directed away from the question "whether usually smoke packet cigarettes, roll-ups or both about equally". Treating these responses as missing values could result in contradicting imputations. For instance, the students who answered "I did once have a puff or two of a cigarette, but I never smoke now" or "I used to smoke sometimes but I never smoke a cigarette now" may be imputed as usually smoking cigarette from a packet, hand-rolled cigarettes, or both, which contradict the former statements made by the students that they might actually smoke a few times in the past but they did not actually smoke currently, in a usual way.

Finally, for all questions generally, the students who did not answer a question, or answered "don't know" as a non-valid option of the question, were recoded as missing.

2.4.3 Collapsing the Levels of Variables

This section concerns the collapsing of levels for several selected variables. When considering the reduction of factor levels, the frequency in each level, the log-odds of each level against drug-trying response variables and whether it is more sensible to combine several levels, should be considered.

One considering factor is the frequency in each level of each variable. If the frequency in a certain level is too low, then it may yield an empty cell in a contingency table with a drug-trying response variable. For example, originally the frequency of respondents of "being excluded" variable (ExclAN1) contained six levels: (1) 0 - No; (2) 1 - Been excluded, but not in the last 12 months; (3) 2 - Once or twice; (4) 3 - 3 to 4 times; (5) 4 - 5 to 10 times and (6) 5 - more than 10 times. The frequency table of the "being excluded" variable is shown in the following table:

Table 2.4.1: Frequency Table of "Being Excluded" Variable

Level	0	1	2	3	4	5	Missing
Frequency	6503	238	287	45	23	5	195

The frequency of respondents in level 5 was considered to be too low, such that in the contingency table against trying anabolic steroids, empty cells occurred, as illustrated by the following table:

Table 2.4.2: Contingency Table of "Being Excluded" Variable against "Tried Anabolic Steroids" Variable

		Being Excluded						
		0	1	2	3	4	5	Missing
Tried anabolic steroids	Yes	20	3	6	1	2	0	116
	No	6430	226	276	42	20	4	2
	Missing	53	9	5	2	1	1	77

As a result, we considered combining level 5 of "Being Excluded" variable with level 3 and level 4 together into a single level, level 3.

Another relevant factor is the log-odds of each level of a variable. If the log-odds against drug-trying response variables are similar, the levels of a variable can be combined. We used the variable that recorded the number of books in a respondent's house, against cannabis as an example. For each level of "Number of books in home" variable, the log-odds of ever trying cannabis could be calculated. The log-odds are shown in the following table:

Table 2.4.3: Log-odds Table of "Number of Books in Home" Variable against "Tried Cannabis" Variable

	Number of books in home					
Levels	None(0)	Very few(1)	11-50(2)	51-100(3)	101-200(4)	>200(5)
Log-odds	-1.6305	-1.8248	-2.2065	-2.3688	-2.6279	-2.5150

From this table, the log-odds for levels 2 to 5 were similar, so these two levels were combined into a single level, level 2.

The frequencies and log-odds for every level of several selected covariates were checked against each drug-trying response variables, before deciding which levels of each of these covariates to be collapsed.

Finally, the levels of variables were checked to determine if these levels were sensible. For several occasions, it might be more sensible if several levels were combined into a single level. For example, for "cigarette smoking status" variable, the following two levels: "I have only ever tried smoking once" and "I used to smoke sometimes but I never smoke a cigarette now", generally meant those smokers used to smoke in the past but they never smoked now. It would be

more sensible to combine these two levels into a single level.

By considering the three criteria mentioned above, several smoking, drinking and drug-related socio-demographic variables were collapsed by the following ways.

2.4.3.1 Variables relating to the Addictive Behaviour of smoking

Regarding the variable about cigarette smoking status, two levels: (1) "not tried" and (2) "ex-smokers", were combined into a single level. Also, two levels: "I sometimes smoke cigarettes now but I don't smoke as many as one a week" and "I usually smoke between one and six cigarettes a week" were combined into a level called "current-light".

When dealing with another variable concerned with the frequency of buying cigarettes from shops in the last year (prior to the survey), three levels: (1) "about once a month"; (2) "two or three times a month" and (3) "once or twice a week", were combined into one level: "occasional".

2.4.3.2 Variables relating to the Addictive Behaviour of drinking

Regarding the variable capturing the frequency of regularly drinking alcohol (AlFreq2), three levels: (1) "once a week"; (2) "twice a week" and (3) "every day or almost every day", were combined into a single level.

When dealing with another variable that captured when a student last had alcohol (AlLast), the two levels of the original variable: (1) "6 months ago or more" and (2) "1 month, but less than 6 months ago", were combined into a level; another pair of levels: (1) "2 weeks, but less than 4 weeks ago" and (2) "1 week, but less than 2 weeks ago" were combined into a level; and the last three levels: (1) "some other time during the last 7 days"; (2) "yesterday" and (3)

"today" were also combined into a single level.

Finally, for a variable that recorded the number of days in last seven days a student drank alcohol, "one to two days" were grouped into a lower level, whilst "three to seven days" were grouped into an upper level.

2.4.3.3 Drug-related Socio-demographic Variables

Considering the variable of how many own age take drugs, the following two levels: (1) "once or twice a week" and (2) "almost every day" were collapsed into a single level.

The four levels of the variable, in respect to the number of books a student had in home: (1) "11 to 50 books"; (2) "51 to 100 books"; (3) "101 to 200 books" and (4) "more than 200 books", were collapsed into a single level.

Also, considering the variable of the frequency of playing truant by a student, three levels: (1) "3 or 4 times"; (2) "5 to 10 times" and (3) "more than 10 times" were collapsed into a single level. A similar collapsing procedure was carried out for the variable in respect of the frequency of being excluded.

2.5 Summary

This chapter has provided a detailed review of survey design, questionnaire design and data source of the Year 2010 Survey. As this research focuses on drug use among young people, as well as for the purposes to reduce the complexity of the original data set of the Year 2010 Survey, the original data set was modified. The modification process of the Year 2010 Survey data set included: (1) proper

recording of the missing data; (2) combining several variables into a single variable, where appropriate, and (3) collapsing factor levels of some variables in the original data set. After the modification of the original data set of the Year 2010 Survey, a cleaner data set, namely "working data set", was obtained, which is more usable for this research. Details of the working data set will be discussed in Chapter 3.

Chapter 3

Exploratory Data Analysis

The Exploratory Data Analysis (EDA) is the best-known work from Tukey (1977), who discussed the need for collecting results of actual data with specific analytic techniques, whilst suggesting the approximation property of actual data on data analysis.

Based on the literature, this chapter describes and summarises the main features of the exploratory data analyses, in respect of the working data set. The purposes to carry out exploratory data analysis of the working data set are to gain more understanding of the properties of the variables in the working data set and the associations among these variables. Section 3.1 provides an overview of the working data set and the variables. In this chapter, we explore the frequencies and percentages for the variables by type in the following sections: (1) the smoking variables in Section 3.1.1; (2) the drinking variables in Section 3.1.2; (3) drug-related socio-demographic variables in Section 3.1.3 and (4) the drug-trying response variables in Section 3.1.4. Section 3.2 further describes the drug-trying response variables. Section 3.3 summarises the pairwise associations among drug-trying response variables and covariates, using contingency tables, log-odds tables, box plots and polychoric correlation plots, where appropriate. The study of the associations and relationships among drug-trying

response variables and covariates (i.e. the smoking, drinking and drug-related socio-demographic variables) has not been carried out in details in the Year 2010 Survey report. It is expected that the aforesaid study will enrich the understanding of drug-trying behaviour among young people in respect of those mentioned covariates.

3.1 Overview of the Working Data Set

After modification of the original data set of the Year 2010 Survey, the working data set of this research contains 68 variables, including 19 smoking variables, 21 drinking variables, 13 drug-related socio-demographic variables and 15 drug-trying response variables. Among these 68 variables, 6 of them are derived variables. Summaries and labels of the variables are presented in Tables A.2.1 to A.2.3 in the Appendix A.2. The sections below provide further details of the variables by sub-types: (1) smoking variables in Section 3.1.1; (2) drinking variables in Section 3.1.2; (3) drug-related socio-demographic variables in Section 3.1.3 and (4) drug-trying response variables in Section 3.1.4.

3.1.1 Smoking Variables

The 19 smoking variables recorded the family attitudes toward smoking (CgFam1), the current cigarette smoking status of respondents (CgStat, CgStat1, Cg7Num, CgNow) and smoking packaging type (CgPk1). In addition, sources of purchasing/obtaining cigarettes (CgGet1, CgGet2, CgGet3, CgGet), the relationship of known smokers (if any) to the students and the estimated proportion of such known smokers (CgPp1, CgEstim) were recorded. Whether the smokers live in the same house as the students (CgWhoSmo, CgWhoHme, CgWho1), the frequency of purchasing cigarettes from shops (CgBuyF1) and obtaining information or lessons about smoking (CgPe1, CgIn1, LsSmk) were also recorded. The combined variable describing the cigarette smoking status (CgStat1) was

created from two separate variables: (1) the average number of cigarettes a student smoked per day in the week prior to the survey (Cg7Num) and (2) the original variable of smoking status (CgStat). We only used this combined variable for this research analysis.

Another combined variable, "smokers in house and where" (CgWho1), was formed from two other separate variables: (1) the number of people living with a student who smoked (CgWhoSmo) and (2) whether people living with a student smoked inside the house (CgWhoHme). A variable describing the usual sources of obtaining cigarettes (CgGet) captured information from three related variables: (1) "number of type of source through shops/ machine/ Internet" (CgGet1); (2) "number of type of source through people" (CgGet2) and (3) number of type of source of being given cigarettes usually by people or other sources" (CgGet3). Tables 3.1.1 and 3.1.2 provide the frequency summaries of the smoking variables respectively, including missing data.

Table 3.1.1: Frequency Table of Smoking Variables (First table)

Variables	Category(Level)	n (%)	Variables	Category(Level)	n (%)
CgFam1	Against(0)	6341 (86.91)	CgStat	Never(0)	5362 (73.49)
	Neutral(1)	598 (8.20)		Tried/Ex-smoker(1)	1264 (17.32)
	For(2)	103 (1.41)		Current-light(2)	385 (5.28)
				Current-moderate to heavy(3)	243 (3.33)
	Missing	254 (3.48)		Missing	42 (0.58)
CgStat1	Never(0)	5358 (73.44)	Cg7Num	0(0)	6528 (89.47)
	Tried/ Ex-smoker(1)	1264 (17.32)		(0,6](1)	325 (4.45)
	Current-light(2)	385 (5.28)		> 6(2)	121 (1.66)
	Current-moderate(3)	95 (1.30)			
	Current-heavy(4)	115 (1.58)			
	Missing	79 (1.08)		Missing	322 (4.41)
CgPk1	None(0)	6626 (90.82)	CgGet1	None(0)	6945 (95.19)
	Packet(1)	273 (3.74)		1(1)	180 (2.47)
	Hand-rolled(2)	96 (1.32)		> 1(2)	94 (1.29)
	Both(3)	219 (3.00)			
	Missing	82 (1.12)		Missing	77 (1.06)
CgGet2	None(0)	6822 (93.50)	CgGet3	None(0)	6626 (90.82)
	Shops only(1)	149 (2.04)		Shops/ people(1)	153 (2.10)
	1(2)	187 (2.56)		1(2)	283 (3.88)
	> 1(3)	61 (0.84)		> 1(3)	157 (2.15)
	Missing	77 (1.06)		Missing	77 (1.06)
CgGet	None(0)	6626 (90.82)	CgPp1	None(0)	1148 (15.73)
	Shops only(1)	81 (1.11)		Other relatives only(1)	958 (13.13)
	People only(2)	52 (0.71)		Friends only(2)	1088 (14.91)
	Given(3)	196 (2.69)		Family members only(3)	606 (8.31)
	Mixture(4)	264 (3.62)		Mixture(4)	2971 (40.72)
	Missing	77 (1.06)		Missing	525 (7.20)
CgWhoSmo	0(0)	4270 (58.53)	CgWhoHme	No(0)	5744 (78.73)
	> 0(1)	2610 (35.77)		Yes(1)	1195 (16.38)
	Missing	416 (5.70)		Missing	357 (4.89)
CgWho1	None(0)	4270 (58.53)	CgBuyF1	Never(0)	6530 (89.50)
	Smoke, outside(1)	1424 (19.52)		Few(1)	199 (2.73)
	Smoke, inside(2)	1174 (16.09)		Occasional(2)	203 (2.78)
				Frequent(3)	55 (0.75)
	Missing	428 (5.87)		Missing	309 (4.24)

Table 3.1.2: Frequency Table of Smoking Variables (Second table)

Variables	Category(Level)	n (%)	Variables	Category(Level)	n (%)
CgEstim	None(0)	1340 (18.37)	CgPe1	None(0)	1340 (18.37)
	Few(1)	3530 (48.38)		Parents, other relatives(1)	4158 (56.99)
	Half(2)	1377 (18.87)		Pros, police(2)	38 (0.52)
	Most, but not all(3)	726 (9.95)		Both(3)	1303 (17.86)
	All(4)	49 (0.67)			
	Missing	274 (3.76)		Missing	457 (6.26)
CgIn1	None(0)	1237 (16.95)	LsSmk	No(0)	1905 (26.11)
	Passive(1)	1298 (17.79)		Yes(1)	4233 (58.02)
	Interactive(2)	309 (4.24)			
	Both(3)	3947 (54.10)			
	Missing	505 (6.92)		Missing	1158 (15.87)
CgNow	No(0)	6504 (89.14)			
	Yes(1)	623 (8.54)			
	Missing	169 (2.32)			

From Tables 3.1.1 and 3.1.2, regarding the CgFam1 variable, a majority of the students' families (86.91 %) were against the students' smoking behaviour, 8.20 % were neutral and 1.41 % were supportive. From the CgStat1 variable, most of the students (90.76 %), including non-smokers and ex-smokers, did not smoke regularly. 5.28 % of the students smoked lightly, 1.30 % of the students smoked moderately, and 1.58 % of the students smoked heavily. Regarding the CgPk1 variable, a few students smoked packet cigarettes (3.74 %), a few students smoked hand-rolled cigarettes (1.32 %) and a few students smoked both (3.00 %). When considering the CgGet variable, a few students obtained cigarettes through shops and people, and were given to them by people (3.62 %). When considering the CgPp1 variable, a majority of the students (77.07 %) reported that either their families or friends or other relatives were smokers. However, when considering the CgWho1 variable, more than half of the students did not have smokers living with them (58.53 %). If the students had smokers living with them, more of these smokers smoked outside their house rather than inside (19.52 % for outside versus 16.09 % for inside). From the CgBuyF1 variable, most smokers never bought any cigarette in the past year (89.50 %), with a few smokers who bought cigarettes occasionally (2.78 %). When considering the

CgEstim variable, a majority of the smokers had a few surrounding smokers (77.87 %), whereas about 29 % of the students had at least half of people of the same age they knew who smoked, and 0.67 % of the smokers reported that all people they knew were smokers. From the CgPe1 and CgIn1 variables, a large proportion of the students received information about smoking from their parents and other relatives (56.99 %), and from both passive and interactive media (54.10 %). Similarly, from the LsSmk variable, more than half of the students had received lessons about smoking (58.02 %). Finally, from the CgNow variable, a majority of the students reported they had never smoked (89.14 %), and less than 10 % of the students reported they had smoked (8.54 %).

In summary, from the Tables 3.1.1 and 3.1.2, a majority of the students reported that: (1) their families were against students' smoking (86.91 %); (2) they did not have smoking habit (90.76 %) and (3) either their families or friends or relatives were smokers (77.07 %). Also, for those students who smoked, 77.87 % of them reported that they knew a few smokers of similar age surrounding them.

3.1.2 Drinking Variables

The 21 drinking variables in the working data set recorded the frequency of drinking alcohol by the students (AlEvr, AlFreq, Al7Day1, AlFreq2), and the last time the students drank alcohol (AlLast). The students' family attitudes towards drinking (AlPar1), places of drinking (AlBnPub, AlUs2), the relationship of known drinkers (if any) to the student (AlUs1), the estimated proportion of known persons who drank (AlEstim) and the number of type of sources and places of purchasing alcohol (AlBuy1, AlBuy2, AlBuy) were recorded. Types of issues happening when drinking (Al4W1), the reason for drinking (AlWhy1) and whether the students had obtained information/education about smoking (LsAlc, AlPe1, AlIn1) were also included.

A variable describing the usual frequency of drinking alcohol (AlFreq2) was derived from two separate variables: (1) the number of days in the week prior to the survey, when alcohol was consumed (Al7Day1) and (2) the frequency of drinking alcohol (AlFreq). A variable describing whether the students usually purchased alcohol themselves or it was obtained via other people (AlBuy) captured information from two related variables: (1) "number of places a student usually purchase alcohol" (AlBuy1) and (2) "number of people from whom a student usually purchase alcohol" (AlBuy2). Another derived variable, "drinkers in house and where" (AlWho1), was combined from two separate variables: (1) "whether people living with the respondent drank inside the house (AlWhoHme)" and (2) "number of people living with respondent who drank (AlWhoDr)".

Tables 3.1.3 and 3.1.4 provide summaries of the drinking variables in terms of frequencies and percentages.

Table 3.1.3: Frequency Table of Drinking Variables (First table)

Variables	Category(Level)	n (%)	Variables	Category(Level)	n (%)
AlEvr	No(0)	3933 (53.91)	AlFreq	Never(0)	3933 (53.91)
	Yes(1)	3271 (44.83)		Ex-drinker(1)	206 (2.82)
				Few a year(2)	1244 (17.05)
				Once a month(3)	557 (7.63)
				Once a fortnight(4)	486 (6.66)
				More than once a fortnight(5)	606 (8.31)
	Missing	92 (1.26)		Missing	264 (3.62)
AlLast	Never(0)	3933 (53.91)	Al7Day1	Did not smoke last week(0)	6075 (83.26)
	up to 1 month ago(1)	1290 (17.68)		1-2 days(1)	790 (10.83)
	4 weeks to 1 week ago(2)	852 (11.68)		3-7 days(2)	146 (2.00)
	During last week(3)	942 (12.91)			
	Missing	279 (3.82)		Missing	285 (3.91)
AlFreq2	Never(0)	3933 (53.91)	AlBnPub	No(0)	5109 (70.02)
	Ex-drinker(1)	206 (2.82)		Yes(1)	1909 (26.17)
	Few a year(2)	1244 (17.05)			
	Once a month(3)	557 (7.63)			
	Current-Light(4)	603 (8.26)			
	Current-Moderate(5)	364 (4.99)			
	Current-Heavy(6)	121 (1.66)			
	Missing	268 (3.67)		Missing	278 (3.81)
AlEstim	None of them(0)	1120 (15.35)	LsAlc	No(0)	1917 (26.27)
	Only a few(1)	2170 (29.74)		Yes(1)	4200 (57.57)
	About half(2)	1574 (21.57)			
	Most, but not all(3)	1966 (26.95)			
	All of them(4)	293 (4.02)			
	Missing	173 (2.37)		Missing	1179 (16.16)
AlPar1	Against(0)	3475 (47.63)	AlBuy1	0 sources(0)	6259 (85.79)
	Middle(1)	3357 (46.01)		1 sources(1)	441 (6.04)
	For(2)	78 (1.07)		2 sources(2)	209 (2.86)
				3 sources or more(3)	70 (0.96)
	Missing	386 (5.29)		Missing	317 (4.34)
AlBuy2	None(0)	5605 (76.82)	AlBuy	None(0)	5605 (76.82)
	From shops(1)	404 (5.54)		Places(1)	404 (5.54)
	1(2)	730 (10.01)		Family members(2)	654 (8.96)
	>1(3)	240 (3.29)		Both(3)	316 (4.33)
	Missing	317 (4.34)		Missing	317 (4.34)

Table 3.1.4: Frequency Table of Drinking Variables (Second table)

Variables	Category(Level)	n (%)	Variables	Category(Level)	n (%)
AlUs1	None(0)	4139 (56.73)	AlUs2	None(0)	4139 (56.73)
	Own(1)	39 (0.53)		Pub(1)	55 (0.75)
	Other people and friends(2)	1134 (15.54)		home/party(2)	1088 (14.91)
	Family members(3)	723 (9.91)		stranger's place/ public outdoor area(3)	623 (8.54)
	Both(4)	979 (13.42)		mixture(4)	1101 (15.09)
	Missing	282 (3.87)		Missing	290 (3.97)
Al4W1	Never in last 4 weeks(0)	6038 (82.76)	AlWhy1	No reasons(0)	494 (6.77)
	Drink, no issue(1)	313 (4.29)		Feel better(1)	239 (3.28)
	Drink, health issue(2)	147 (2.01)		Socialise(2)	690 (9.46)
	Drink, aggressive and other issue(3)	139 (1.91)		Both(3)	5550 (76.07)
	Drink, both(4)	240 (3.29)			
	Missing	419 (5.74)		Missing	323 (4.43)
AlWhoDr	0(0)	1351 (18.52)	AlWhoHme	No(0)	2424 (33.22)
	>0(1)	5420 (74.29)		Yes(1)	4458 (61.10)
	Missing	525 (7.20)		Missing	414 (5.67)
AlWho1	None(0)	1351 (18.52)	AlPe1	None(0)	1367 (18.74)
	Drink, outside(1)	988 (13.54)		Parents, other relatives(1)	4092 (56.09)
	Drink, inside(2)	4416 (60.53)		Pros, police(2)	32 (0.44)
				Both(3)	1309 (17.94)
	Missing	541 (7.42)		Missing	496 (6.80)
AlIn1	None(0)	1452 (19.90)			
	Passive Media(1)	1407 (19.28)			
	Interactive Media(2)	259 (3.55)			
	Both(3)	3632 (49.78)			
	Missing	546 (7.48)			

Regarding the CgStat1 variable in Table 3.1.1 and the AlEvr variable in Table 3.1.3, drinkers and non-drinkers were much more evenly distributed than smokers and non-smokers (44.83 % and 53.91 % compared to 25.48 % and 73.44 % respectively). From the AlFreq and AlFreq2 variables, while 17.05 % of the students drank a few times a year, 8.26 % of the students drank every fortnight (current-light), and 1.66 % of the students drank at least three days in the previous week (current-heavy). These figures were further augmented by the figures

from the AlLast variable that 17.68 % of the students drank alcohol up to the previous month and 12.91 % of the students drank alcohol during the previous week prior to the survey.

From the AlBnPub variable in Table 3.1.3, most of the students (70.02 %) had not been to the pub, but from the AlEstim variable, 82.28 % of the students were surrounded by drinkers. From the AlPar1 variable, the majority of the family members of the students (93.64 %) were either against drinking or neutral to drinking alcohol. On the other hand, from the AlBuy variable, 8.96 % of the students obtained alcohol from their family members, 5.54 % of the students obtained alcohol from various places such as supermarkets and 4.33 % of the students obtained alcohol from both these source types. Referring to Table 3.1.4, from the AlUs1 variable, very few students (0.53 %) drank alcohol on their own, 15.54 % of them drank alcohol with other people and friends, whereas 9.91 % of them drank alcohol with family members, and 13.42 % of the students drank alcohol with both groups of people. Consequently, from AlUs2 variable, 14.91 % of the students drank alcohol at home or at a party, 8.54 % of the students drank in other places and 15.09 % of the students drank alcohol at home and/or at a party and/or in other places. Finally, from the AlWho1 variable, more than half of the students (60.53 %) had drinkers at home. These figures potentially reflected that despite unfavourable opinions from families about drinking alcohol, plenty of the students consumed alcohol at home with their friends and family members, and they were surrounded by drinkers at home.

Additionally, from the Al4W1 variable, among the small proportion of students who reported that they had drunk in the past four weeks, more students had issues associated with alcohol than those who did not (7.21 % versus 4.29 %). From the AlWhy1 variable, most of these students (88.81 %) stated the reasons that people drank to feel better and/or to socialise with other people. On the

other hand, from the LsAlc variable, more than half of the students (57.57 %) received lessons about drinking. In addition, from the AlPe1 variable, more than half of the students (56.09 %) received information about drinking from their parents and other relatives, and from the AlIn1 variable, about half of the students (49.78 %) received information about drinking from both passive and interactive media.

In summary, from the Tables 3.1.3 and 3.1.4, about half of the students (53.91 %) reported that they did not drink. However, when compared with the students who smoked, there were more students who drank (44.83 %) than who smoked (25.48 %). Majority of the students' family members (93.64 %) were either against drinking or neutral to drinking alcohol. Also, 82.28 % of the students reported that they were surrounded by drinkers. For those students reported that they were drinkers, 60.53 % of them reported that they had drinkers at home and a majority of them (88.81 %) drank for feeling better and/or socialisation reasons.

3.1.3 Drug-Related Socio-demographic Variables

Demographic information relating to age (Age) and gender (Gender) of the students were available, as were information regarding drug knowledge (LsDrg, DgPe1, DgIn1) and the estimated proportion of peer (own age) drug use (DgEstim). In addition, the information on truancy (TruantN, Truant1), exclusion from school (ExclAN1, ExclA1), number of books in the home (Books1), whether a student had enrolled in a free school meal scheme (FSM1) and the students' Strategic Health Authority regions (SHA) were included.

Table 3.1.5 provides summaries of the drug-related socio-demographic variables, in terms of frequency and percentages.

Table 3.1.5: Frequency Table of Drug-related Socio-demographic Variables

Variables	Category(Level)	n (%)	Variables	Category(Level)	n (%)
DgPe1	None(0)	2091 (28.66)	DgIn1	None(0)	1687 (23.12)
	Parents, other relatives(1)	3154 (43.23)		Passive Media(1)	1115 (15.28)
	Pros, police(2)	81 (1.11)		Interactive Media(2)	489 (6.70)
	Both(3)	1431 (19.61)		Both(3)	3466 (47.51)
	Missing	539 (7.39)		Missing	539 (7.39)
DgEstim	None(0)	3170 (43.45)	Books1	None(0)	292 (4.00)
	Only a few(1)	3272 (44.85)		Very few 1 to 10(1)	943 (12.92)
	About half(2)	484 (6.63)		Enough to fill 1 shelf and more(2)	5776 (79.17)
	Most to all(3)	186 (2.55)			
	Missing	184 (2.52)		Missing	285 (3.91)
LsDrg	No(0)	1819 (24.93)	Age	11 years old(0)	1154 (15.82)
	Yes(1)	4238 (58.09)		12 years old(1)	1502 (20.59)
				13 years old(2)	1486 (20.37)
				14 years old(3)	1468 (20.12)
				15 years old(4)	1686 (23.11)
	Missing	1239 (16.98)		Missing	0 (0)
Gender	Boy(0)	3688 (50.55)	FSM1	No(0)	6058 (83.03)
	Girl(1)	3608 (49.45)		Yes(1)	1001 (13.72)
	Missing	0 (0)		Missing	237 (3.25)
Truant1	No(0)	6181 (84.72)	TruantN	No(0)	6181 (84.72)
	Yes(1)	879 (12.05)		Played truant, not in last 12 months(1)	231 (3.17)
				Once/ twice(2)	401 (5.50)
				>= 3 times(3)	213 (2.92)
	Missing	236 (3.23)		Missing	270 (3.70)
ExclA1	No(0)	6503 (89.13)	ExclAN1	No(0)	6503 (89.13)
	Yes(1)	606 (8.31)		Excluded, not in last 12 months(1)	238 (3.26)
				1-2 times(2)	287 (3.93)
				>= 3 times(3)	73 (1.00)
	Missing	187 (2.56)		Missing	195 (2.67)
SHA	North East(0)	699 (9.58)	SHA	East(5)	756 (10.36)
	North West/ Merseyside(1)	710 (9.73)		London(6)	491 (6.73)
	Yorkshire and the Humber(2)	503 (6.89)		South East Coast(7)	769 (10.54)
	East Midlands(3)	814 (11.16)		South Central(8)	842 (11.54)
	West Midlands(4)	946 (12.97)		South West(9)	766 (10.50)
	Missing	0 (0)			

Regarding the Gender variable in Table 3.1.5, the percentages of boys and girls were similar. Regarding the Age variable, except the percentage of 11 years old students, which was 15.82 %, and the percentage of 15 years old students, which was 23.11 %, all other levels yielded percentages of approximately 20 %. Considering the DgPe1 and DgIn1 variables, 43.23 % of the students received information about drugs from their parents and other relatives, and about 47.51 % of the students received information from both passive and interactive media, such as FRANK and the Internet. Moreover, when considering the LsDrg variable, most students (58.09 %) had lessons about drugs in the last 12 months. These findings were similar to those variables related to smoking and drinking education and information (CgPe1, CgIn1, LsSmk, AlPe1, AlIn1 and LsAlc respectively). In addition, from the DgEstim variable, nearly half of the students (44.85 %) knew only a few persons who had tried drugs, and from the Books1 variable, most students (79.17 %) possessed books that were filled at least one bookshelf.

From the FSM1 variable, a majority of the students in this survey (83.03 %) were not enrolled in a free school meal scheme, this might suggest that the families of most students were not in economic difficulties, according to Hobbs and Vignoles (2007). In addition, from the Truant1 variable, 84.72 % of the students had not played truant and from the TruantN variable, 3.17 % of the students had played truant more than 12 months ago. However, several of the students (5.50 %) had truanted once or twice in the year, and 2.92 % of the students had truanted at least three in the year. From the ExclA1 variable, most students (89.13 %) had not been excluded from school, and from the ExclAN1 variable, only 3.93 % of the students had been excluded once or twice in that year.

3.1.4 Drug-trying Response Variables

The drug-trying response variables in this research analysis were the 15 drugs, containing information about whether a student had ever tried a specific drug (yes/no). The 15 drugs were: (1) cannabis; (2) heroin; (3) cocaine; (4) magic mushrooms; (5) crack; (6) methadone; (7) ecstasy; (8) amphetamines; (9) LSD; (10) poppers; (11) ketamine; (12) anabolic steroids; (13) gas; (14) other drugs and (15) tranquillisers. Table 3.1.6 provides the frequency and percentages summary of drug-trying response variables.

Table 3.1.6: Frequency Table of Drug-trying Response Variables

Variables	Category (Level)	n (%)	Variables	Category (Level)	n (%)
DgTdCan1 (Cannabis)	No(0)	6485 (88.88)	DgTdHer1 (Heroin)	No(0)	7104 (97.37)
	Yes(1)	661 (9.06)		Yes(1)	36 (0.49)
	Missing	150 (2.06)		Missing	156 (2.14)
DgTdCok1 (Cocaine)	No(0)	7060 (96.77)	DgTdMsh1 (Magic Mushrooms)	No(0)	7031 (96.37)
	Yes(1)	87 (1.19)		Yes(1)	109 (1.49)
	Missing	149 (2.04)		Missing	156 (2.14)
DgTdCrk1 (Crack)	No(0)	7105 (97.38)	DgTdMth1 (Methadone)	No(0)	7085 (97.11)
	Yes(1)	45 (0.62)		Yes(1)	52 (0.71)
	Missing	146 (2.00)		Missing	159 (2.18)
DgTdEcs1 (Ecstasy)	No(0)	7058 (96.74)	DgTdAmp1 (Amphetamines)	No(0)	7056 (96.71)
	Yes(1)	80 (1.10)		Yes(1)	67 (0.92)
	Missing	158 (2.17)		Missing	173 (2.37)
DgTdLSD1 (LSD)	No(0)	7113 (97.49)	DgTdPop1 (Poppers)	No(0)	6979 (95.66)
	Yes(1)	42 (0.58)		Yes(1)	164 (2.25)
	Missing	141 (1.93)		Missing	153 (2.10)
DgTdKet1 (Ketamine)	No(0)	7119 (97.57)	DgTdAna1 (Anabolic Steroids)	No(0)	7114 (97.51)
	Yes(1)	43 (0.59)		Yes(1)	34 (0.47)
	Missing	134 (1.84)		Missing	148 (2.03)
DgTdGas1 (Gas)	No(0)	6569 (90.04)	DgTdOth1 (Other Drugs)	No(0)	7120 (97.59)
	Yes(1)	590 (8.09)		Yes(1)	33 (0.45)
	Missing	137 (1.88)		Missing	143 (1.96)
DgTdTrn1 (Tranquillisers)	No(0)	7129 (97.71)			
	Yes(1)	32 (0.44)			
	Missing	135 (1.85)			

From Table 3.1.6, cannabis was the most commonly used drugs amongst the students with 9.06 % of the students having tried it. This was followed by gas and poppers, with 8.09 % of the students reported that they had tried gas and 2.25 % of the students reported that they had tried poppers. The least-used drug group was tranquillisers, which was only used by 0.44 % of the students.

In this research, further investigation of the demographic properties of drug-trying response variables was carried out and further exploration among these variables will be discussed in Section 3.2.

3.2 Further Exploration among Drug-trying Response Variables

In this stage of the exploration among drug-trying response variables, the total number of drugs each student had tried was investigated. The rationale was to gauge the potential level of drug-trying behaviour of the students. The frequency table of number of drugs tried by the students, based on observed data, is provided in Table 3.2.1.

Table 3.2.1: Frequency Table of Number of Drugs Tried by Students

# Drugs Tried	0	1	2	3	4	5	6
Frequency	6094	821	199	71	45	25	9
Percent	83.53%	11.25%	2.73%	0.97%	0.62%	0.34%	0.12%
# Drugs Tried	7	8	9	10	11	12	13
Frequency	10	6	6	4	2	2	2
Percent	0.14%	0.08%	0.08%	0.05%	0.03%	0.03%	0.03%

From Table 3.2.1, although majority of the students (83.53 %) reported they tried no drugs, 821 students (11.25 %) reported they had tried one drug, 381 students (5.22 %) reported that they had tried at least two drugs, including 199 that had tried two drugs, 71 that had tried three drugs, 45 that had tried four drugs, 25 that had tried five drugs, and 41 that had tried at least six drugs. The above result of 16.47 % of 7,296 students participated in the Year 2010 Survey reported that they had tried different drugs indicates that there may exist a high number of young people in England, estimated to be 450,000 young people by basing on the estimated total number of 3 million of boys and girls aged 11 to 15 in England by the Year 2010 Survey report (Fuller et al., 2011), who have taken drugs. This

further supports the prior research finding of a sustained prevalence of drug use among young people in British society and that further research effort should be continuously devoted to address the drug use problem.

3.3 Pairwise Associations between Drug-trying Response Variables and Covariates

The Year 2010 Survey report did not study the associations among drug-trying response variables and covariates (i.e. the smoking, drinking and drug-related socio-demographic variables) in detail. To understand more about the drug-trying behaviour of young people in respect of those covariates, in this section, pairwise associations among drug-trying response variables and the smoking, drinking and drug-related socio-demographic variables were depicted by percentage contingency tables, box plots (for continuous variables) and polychoric correlation plots.

3.3.1 Percentage Contingency Tables among Covariates and Drug-trying Response Variables

In this section, we examine the pairwise associations between the categorical covariates and the drug-trying response variables. Percentage contingency tables were adopted to investigate such pairwise associations. In the percentage tabulates, drug-trying response variable for drug A, on the horizontal x-axis, was compared to covariate B, on the vertical y-axis, to investigate the percentage of students who had tried drug A against each factor level of covariate B (i.e. given a factor level of covariate B, what was the percentage of students who had tried drug A). The purpose of using percentage tabulates was to illustrate the increase in the percentage of students who had tried drug A, when we set the factor level of covariate B from lower level to higher level. A positive associa-

tion can be observed from the percentage tabulates if the percentage increases corresponding to rises in factor levels of the covariate.

Several covariates were selected to represent smoking, drinking and drug-related socio-demographic variables, in order to produce percentage tabulates in respect of these selected covariates and drug-trying response variables. These selected covariates are listed in Table 3.3.1 as follows:

Table 3.3.1: Table of Selected Covariates for Depiction

Category	Selected Covariate
Smoking	CgFam1, CgStat1, CgWho1, CgBuyF1
Drinking	AlFreq2, AlPar1, AlBuy, AlWho1
Drug-related Socio-demographic	DgEstim, Age, Gender, FSM1, TruantN, ExclAN1

These covariates were selected because they were the most informative variables that might predict drug-trying behaviour within their own variable group (as discussed in Sections 3.1.1, 3.1.2 and 3.1.3). For the smoking variables, the CgFam1 variable represented the responses of the students' families to smoking, whereas the CgStat1 and the CgBuyF1 variables represented the cigarette smoking status and frequency of purchasing cigarettes of the students respectively. Also, the CgWho1 variable represented the number of smokers that the students had in their houses. For the drinking variables, the AlFreq2 variable represented the students' frequency of drinking alcohol, whereas the AlBuy variable represented the students' sources of obtaining alcoholic drinks. The AlPar1 variable represented the students' family responses to drinking, and the AlWho1 variable represented the number of alcohol drinkers that the students had in their houses. For drug-related socio-demographic variables, the DgEstim variable represented the proportion of drug takers around the students. The Age and Gender variables were demographic variables that were usually related to drug-trying behaviour. The FSM1 variable represented the economic status of the students' families, whereas the TruantN and ExclAN1 variables repre-

sented the behaviour of the students in playing truancy and being excluded from schools respectively. The description of these variables in this section can be referred to Appendix A.2.

In Tables 3.3.2 and 3.3.3 the percentage tabulates portrayed the conditional percentages of the students who had tried a specific drug, which were listed along the x-axis, given a factor level of a covariate listed along the y-axis.

Table 3.3.2: Percentage Contingency Table of Drug-trying Response Variable against Smoking, Drinking and Drug-related Socio-demographic Variables (First Table). Every cell represents the conditional percentage of students who had tried a specific drug, given a factor level of a covariate listed along the y-axis.

Variable	Level	Can.	Heroin	Cocaine	Mush.	Crack	Meth.	Ecstasy	Amph.	LSD	Poppers	Ket.	Steroids	Gas	Other	Tranq.
CgFam1	0	7.98	0.40	0.93	1.19	0.50	0.50	0.85	0.69	0.42	1.79	0.40	0.43	7.93	0.38	0.37
	1	12.37	0.52	2.05	2.07	0.52	1.03	1.38	1.91	1.03	3.44	1.20	0.34	8.97	0.86	0.68
	2	48.51	6.06	12.12	15.00	10.10	12.00	14.14	11.88	6.93	18.00	6.00	3.00	19.00	2.02	5.00
CgStat1	0	0.74	0.15	0.27	0.29	0.19	0.17	0.19	0.15	0.08	0.46	0.15	0.25	6.31	0.08	0.17
	1	19.45	0.81	1.78	1.87	1.29	0.97	1.14	0.89	1.13	3.25	0.48	0.89	11.44	0.57	0.56
	2	52.89	1.60	3.99	8.27	2.12	2.41	5.35	4.24	1.33	11.73	4.24	1.33	17.68	1.85	1.06
	3	64.21	3.23	3.16	10.64	3.16	7.45	8.51	7.45	3.26	10.64	5.26	1.05	9.47	1.05	3.23
	4	85.96	6.36	27.52	20.18	5.50	12.50	20.35	17.70	11.50	34.82	6.19	3.67	28.57	12.39	5.36
CgWho1	0	5.98	0.19	0.80	0.92	0.52	0.52	0.73	0.59	0.26	1.56	0.47	0.21	8.15	0.35	0.21
	1	14.31	0.64	1.27	2.13	0.85	0.85	1.35	1.21	0.85	3.34	0.92	0.64	9.06	0.57	0.92
	2	15.70	1.21	2.50	2.94	0.69	1.04	1.83	1.65	1.12	3.64	0.43	0.95	8.34	0.69	0.69
CgBuyF1	0	5.69	0.28	0.64	0.85	0.43	0.36	0.53	0.42	0.26	1.08	0.23	0.36	7.68	0.26	0.25
	1	46.67	2.08	6.70	6.81	3.63	3.12	6.77	7.18	4.15	17.71	4.66	1.04	19.59	1.05	3.65
	2	69.50	3.65	9.09	13.64	3.09	7.18	10.77	7.58	5.08	18.37	8.12	2.05	14.07	5.03	2.05
	3	70.59	11.54	23.53	19.23	7.69	11.54	21.57	19.61	13.21	39.22	3.92	9.62	21.15	5.77	7.69
AlFreq2	0	0.77	0.15	0.34	0.31	0.26	0.23	0.21	0.18	0.18	0.28	0.15	0.21	5.34	0.10	0.26
	1	9.00	0.99	2.00	0.99	1.51	0.49	1.00	0.50	1.00	0.99	1.00	0.00	8.50	1.48	0.99
	2	5.55	0.25	0.49	0.74	0.41	0.08	0.25	0.08	0.16	0.82	0.41	0.49	9.48	0.08	0.33
	3	18.69	0.55	1.10	1.48	0.19	1.10	1.83	1.28	0.18	2.94	0.56	0.92	11.01	0.92	0.18
	4	31.76	1.18	2.36	4.22	1.18	1.53	2.03	2.03	1.35	6.27	1.51	0.51	14.45	1.17	0.84
	5	45.00	1.69	5.92	8.12	2.25	4.49	6.78	5.07	2.25	15.25	3.66	1.69	16.43	1.68	1.41
	6	56.67	7.63	17.65	15.00	9.17	4.20	15.83	14.29	10.00	21.67	2.50	5.00	21.67	4.27	1.67
AlPar1	0	4.18	0.41	0.73	0.64	0.47	0.38	0.35	0.41	0.23	0.90	0.26	0.29	5.96	0.23	0.26
	1	13.73	0.48	1.63	2.33	0.75	0.88	1.66	1.36	0.85	3.32	0.79	0.64	10.50	0.70	0.48
	2	41.56	5.33	8.33	6.85	5.26	8.22	9.21	6.67	5.26	18.67	5.33	4.11	10.39	1.32	3.95

Table 3.3.3: Percentage Contingency Table of Drug-trying Response Variable against Smoking, Drinking and Drug-related Socio-demographic Variables (Second Table) Every cell represents the conditional percentage of students who had tried a specific drug, given a factor level of a covariate listed along the y-axis.

Variable	Level	Can.	Heroin	Cocaine	Mush.	Crack	Meth.	Ecstasy	Amph.	LSD	Poppers	Ket.	Steroids	Gas	Other	Tranq.
AlBuy	0	2.93	0.25	0.47	0.53	0.36	0.24	0.31	0.22	0.27	0.58	0.23	0.29	6.77	0.16	0.32
	1	39.70	2.02	6.06	5.29	2.27	2.77	6.15	5.79	2.78	12.31	2.51	2.02	13.10	1.53	1.26
	2	25.97	0.47	2.18	4.52	0.93	1.87	2.33	2.34	0.62	5.00	1.40	0.62	11.99	1.71	0.16
	3	48.39	3.64	6.95	8.94	2.97	4.01	7.17	4.87	3.57	14.24	3.30	1.98	20.39	1.95	1.63
AlWho1	0	4.89	0.30	0.52	0.97	0.45	0.45	0.53	0.15	0.22	0.52	0.15	0.22	6.30	0.22	0.22
	1	5.83	0.41	0.51	0.82	0.51	0.51	0.61	0.72	0.31	1.13	0.51	0.41	5.01	0.00	0.20
	2	11.53	0.53	1.58	1.85	0.69	0.76	1.35	1.17	0.69	3.09	0.71	0.48	9.79	0.64	0.55
DgEstim	0	1.98	0.13	0.22	0.38	0.25	0.22	0.03	0.13	0.06	0.48	0.13	0.26	3.63	0.13	0.22
	1	11.18	0.43	1.29	1.48	0.55	0.49	1.17	0.83	0.52	2.28	0.43	0.37	9.71	0.34	0.28
	2	31.17	1.27	3.35	4.84	1.90	2.75	4.24	4.67	2.11	9.28	3.55	1.26	23.01	2.94	1.88
	3	45.36	6.63	12.15	14.29	5.56	8.38	10.38	7.14	7.07	16.57	4.42	3.89	24.46	2.19	3.87
Age	0	0.27	0.35	0.53	0.35	0.09	0.09	0.36	0.27	0.27	0.53	0.18	0.18	6.76	0.09	0.35
	1	1.15	0.27	0.27	0.41	0.41	0.27	0.34	0.14	0.20	0.34	0.14	0.41	6.88	0.07	0.27
	2	5.52	0.69	0.75	1.65	0.68	0.55	0.55	0.62	0.62	1.51	0.34	0.41	7.87	0.34	0.48
	3	10.53	0.56	1.11	1.81	0.90	0.56	1.25	1.18	0.62	2.71	0.42	0.56	9.38	0.56	0.62
	4	24.56	0.60	3.02	2.96	0.91	1.87	2.72	2.17	1.08	5.57	1.68	0.72	9.79	1.08	0.48
Gender	0	10.02	0.47	1.20	1.62	0.81	0.70	1.06	0.92	0.67	2.20	0.61	0.64	7.22	0.39	0.56
	1	8.47	0.53	1.23	1.43	0.45	0.76	1.18	0.96	0.51	2.39	0.59	0.31	9.27	0.53	0.34
FSM1	0	9.12	0.47	1.17	1.45	0.63	0.65	1.14	1.00	0.57	2.34	0.55	0.47	8.13	0.50	0.40
	1	10.14	0.71	1.41	1.92	0.61	1.11	1.02	0.51	0.50	2.22	1.01	0.40	9.05	0.10	0.70
TruantN	0	5.56	0.24	0.52	0.91	0.34	0.33	0.57	0.44	0.34	1.16	0.33	0.26	6.88	0.18	0.23
	1	29.44	1.77	6.52	5.73	3.06	2.20	3.93	4.35	2.60	7.42	1.75	1.75	16.16	1.75	0.87
	2	33.16	1.02	3.05	3.80	1.52	2.78	3.05	2.77	1.01	8.95	1.52	0.76	17.22	2.03	1.27
	3	48.82	6.28	12.68	11.16	4.81	6.76	11.00	8.74	4.29	16.83	5.74	3.85	22.38	4.37	4.76
ExclAN1	0	7.03	0.33	0.82	1.17	0.42	0.50	0.78	0.61	0.40	1.57	0.39	0.31	7.72	0.23	0.28
	1	27.04	1.30	4.74	4.26	3.40	2.56	4.74	2.16	1.72	6.52	2.58	1.31	15.68	1.72	2.16
	2	35.69	2.14	5.02	4.29	2.14	2.50	3.91	5.38	1.77	11.43	2.83	2.13	12.01	3.18	1.77
	3	47.89	8.70	10.14	11.43	4.35	7.14	10.00	8.45	8.45	17.14	4.35	4.35	19.72	4.29	4.29

From Tables 3.3.2 and 3.3.3, a general trend was observed that in most cases the higher the levels of a selected smoking, drinking or drug-related socio-demographic variable, the more likely that a student had tried a drug. The percentage tabulates of CgStat1, CgBuyF1 and AlFreq2 variables indicated that generally the students' heavier smoking or drinking habits were linked to their increased drug-trying behaviour. The percentage tabulates of DgEstim showed peer influence on the students' drug-trying behaviour that the more people of the same age who had tried drugs, the higher the likelihood that the students would try drugs. These two phenomena were particularly obvious for the six drugs: (1) cannabis; (2) poppers; (3) cocaine; (4) ecstasy; (5) magic mushrooms and (6) gas. In addition, percentage tabulates of CgFam1 and AlPar1 variables indicated that for almost every drug, when the students' families were inclined to support the students' smoking or drinking behaviour, such students became more likely to try drugs. The percentage tabulates of the CgWho1 variable generally illustrated that the students, who had smokers living with them and smoking inside their houses, were more likely to try drugs than those who have smokers smoking outside their houses. In addition, the latter group of the students was more likely to try drugs than those who had no smokers living with them. Similar findings were found on AlWho1 variable. From the percentage tabulates of the AlBuy variable, it was observed that purchasing alcohol from shops influenced more students to try a drug than purchasing alcohol from family members. This finding may imply a relationship of places where alcohol and drugs could be bought.

However, there was no significant gender difference of trying types of drugs except that girls apparently used more gas than boys (9.27 % for girls and 7.22 % for boys), whereas boys apparently used more cannabis than girls (10.02 % for boys and 8.47 % for girls). For the free school meal, represented by FSM1 variable as a proxy of the economic status of the students' families, it was observed

that the students involved in the free school meal scheme were more likely to try cannabis, heroin, cocaine, magic mushrooms, methadone, ketamine, gas and tranquillisers. This might imply a relationship of economic status of the students' families and students' drug-trying behaviour.

From the percentage tabulates of both TruantN and ExclAN1 variables, a general pattern was observed that the percentages of the students trying drugs increased when the frequencies of the students playing truant or being excluded from school increased. This finding implies positive associations between both TruantN and ExclAN1 variables and drug-trying response variables.

Additionally, from the percentage tabulates of the Age variable, it was observed that generally, older students were increasingly likely to try drugs. The positive correlation between Age variable and drug-trying response variable was particularly strong in respect of the drugs: (1) cannabis; (2) gas; (3) poppers; (4) cocaine and (5) magic mushrooms.

3.3.2 Empty Cell Problem

The empty cell problem, which means zero cell count for a combination of factor levels from both categorical variables, existed in some combinations of covariates and drug-trying response variables, such as DgPe1 and heroin, and DePe1 and tranquillisers. The contingency tabulates of DgPe1 against heroin and DgPe1 against tranquillisers, shown in Table 3.3.4, were used as examples to assist in explaining the empty cell problem.

Table 3.3.4: Contingency Tablulates of DgPe1 against Heroin and Tranquillisers

Drug-trying Response		Heroin		Tranquillisers	
Variable	Level	No	Tried	No	Tried
DgPe1	None (0)	2052	12	2063	8
	Parents, other relatives (1)	3117	14	3123	8
	Pros, police (2)	80	0	81	0
	Both (3)	1404	8	1407	12

These cross-tabulations in Table 3.3.4 showed no students who received information about drugs from the professionals and the police, had tried heroin or tranquillisers. Empty cells cause problems that lead to undefined likelihood estimates, since the log of zero is undefined. It should also be noted that sparse data were detected as well if there were very low frequencies in several frequency cells. However, this type of sparse data would not lead to singularities when fitting logistic regression models. One example of such sparse data with positive frequencies was the cell representing the frequency of the students having tried heroin and obtained information from professionals and the police (Level 2 of DgPe1 Variable).

3.3.3 Box Plots for Continuous Variables

To investigate the relationships between the continuous variables and drug-trying response variables, box plots were adopted as well. In the working data set, there were three continuous variables: Cg7Num, CgWhoSmo and Age. For each continuous variable, we plotted fifteen box plots, with each box plot showed each continuous variable against a single binary drug-trying response variable. In addition, we plotted a box plot, in which the AlEvr variable, under the label 'Alcohol', was analysed against each of the three continuous variables, in order to investigate the relationships between the three continuous variables and this drinking variable. These box plots are plotted, clustered and presented in Figures 3.1 and 3.2.

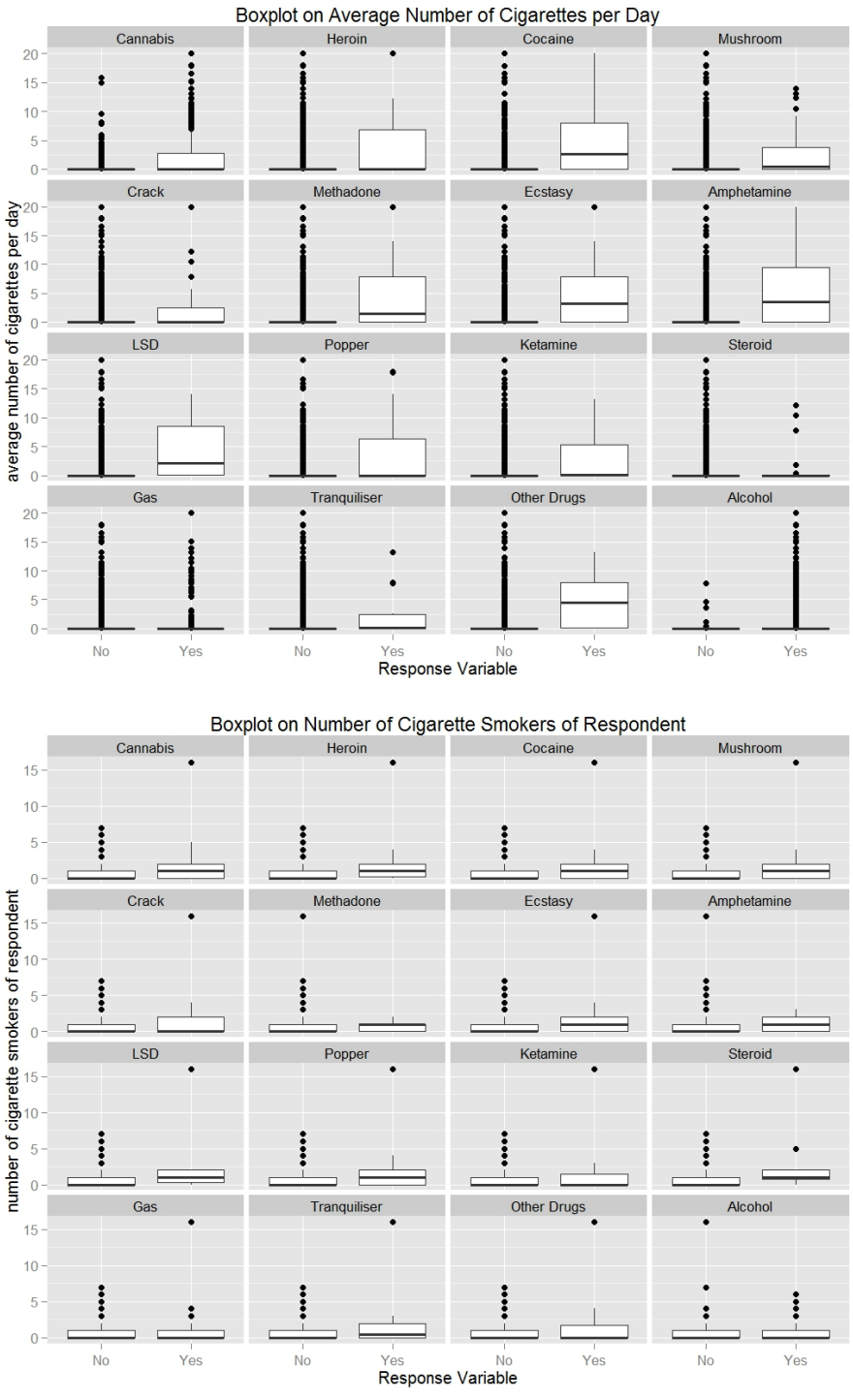

Figure 3.1: Box Plots for the Average Number of Cigarettes per day (Cg7Num) and Number of Cigarette Smokers of Respondent (CgWhoSmo) Covariates against Drug-trying Response Variables and "Alcohol" Variables

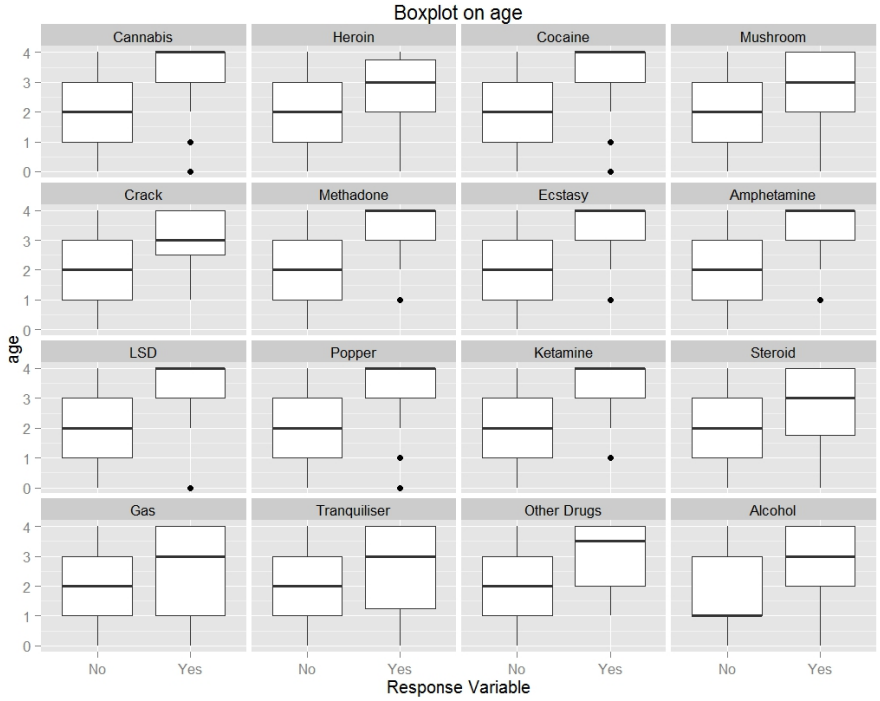

Figure 3.2: Box Plots for Age against Drug-trying Response Variables and "Alcohol" Variables

From the box plots in Figures 3.1 and 3.2, the Cg7Num variable exhibited positive associations with most drug-trying response variables (except anabolic steroids and gas). When examining the medians between drug-triers and non drug-triers for cocaine, methadone, ecstasy, amphetamines, LSD and other drugs, positive associations between the average number of cigarettes per day in the previous week and these drug-trying response variables were observed. This finding reflects drug-triers' tendency to smoke more cigarettes on average in the previous week. No apparent observations were made from box plots with CgWhoSmo variable. Finally, Age variable was found to be significantly related to alcohol covariate, as well as cannabis, cocaine, crack, methadone, ecstasy, amphetamines, LSD, poppers and ketamine, implying that drug-triers and drinkers in this survey are usually older students.

3.3.4 Polychoric Correlation Plots

In this section, polychoric correlation plots were adopted to investigate the relationships between variables in the working data set. These plots were adopted to illustrate polychoric associations among a large number of variables at a glance. In each correlation plot, listwise comparison between variable pairs was adopted to generate comparable correlation values across every correlation plot.

Polychoric correlation (Drasgow, 1986) is a method of measuring the correlation between two ordinal or continuous variables. Polychoric correlation is scaled between -1 and 1, and can be applied to continuous, ordinal and binary variables. Nominal variables are broken down into separate binary factors that correspond to each level in these variables. Variables in a plot are ordered according to their aggregate magnitude in polychoric correlations. The resulting polychoric correlation plots are presented in Figures 3.3 to 3.5.

Firstly, Figure 3.3 illustrated that most smoking variables yielded strong positive correlations with drug-trying response variables, and that these drug-trying response variables were strongly and positively correlated with each other, particularly for cannabis, cocaine and amphetamines. It was observed that in particular, CgFam1, CgGet, CgPk1, CgEstim, CgBuyF1, CgGet1, CgGet2, CgGet3, CgStat1, CgStat, CgNow and Cg7Num variables were highly positively correlated with drug-trying response variables. These findings implied that, family's attitude to smoking, cigarette smoking status, number of cigarettes smoked, frequency of purchasing cigarettes, sources of obtaining cigarettes, as well as the proportion of people a student knows who smoke, are all positively associated with drug-trying response variables. These smoking variables were also found highly correlated with each other as well.

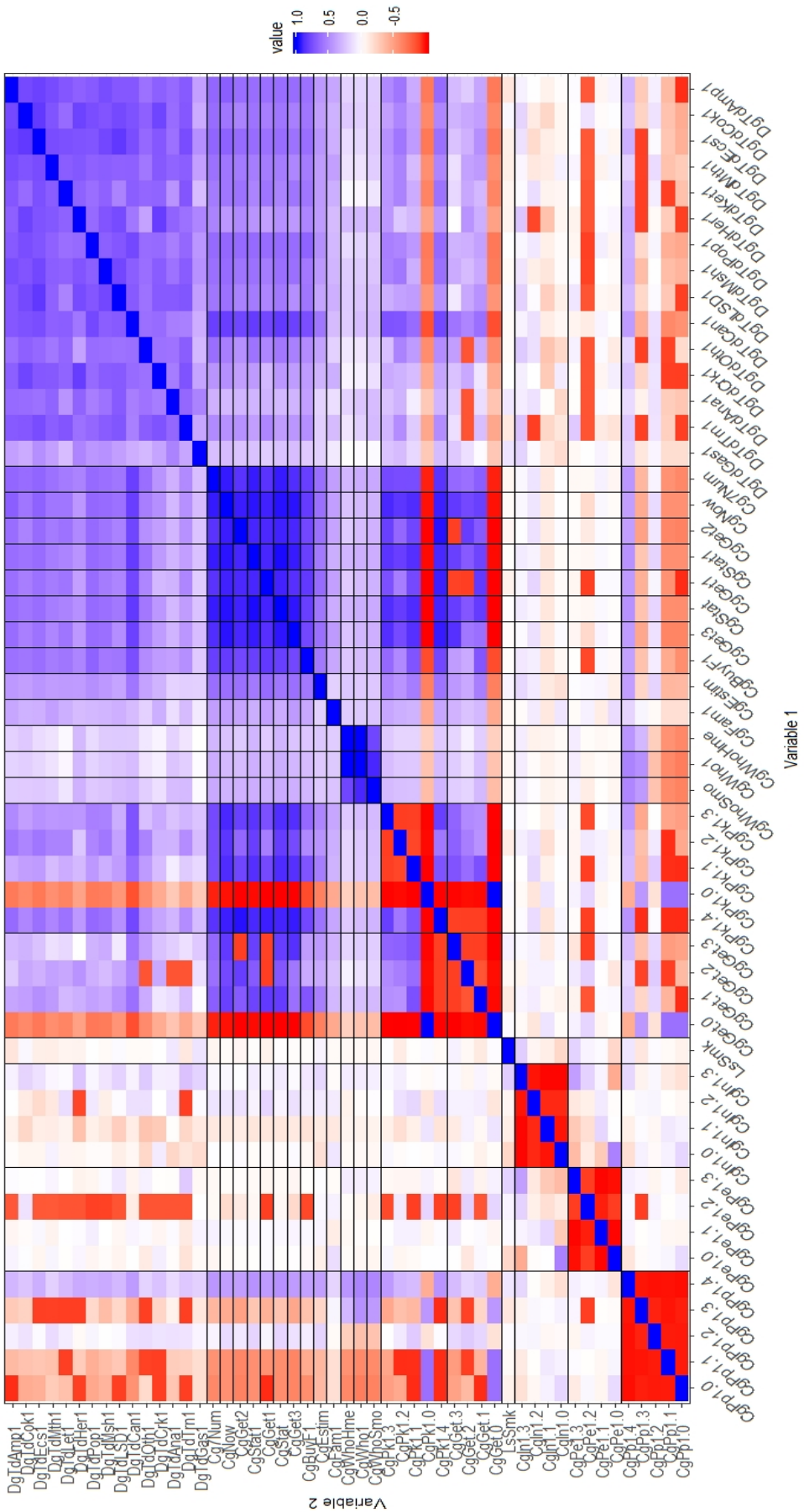

Figure 3.3: Polychoric Correlation Plot among the Smoking and Drug-trying Response Variables. Smoking variables were indicated as from CgPp1 to Cg7Num, at lower rows and columns at the left, drug-trying response variables were indicated as from DgTdGas1 to DgTdAmp1, at upper rows and columns at the right. Red indicated negative correlation and blue indicated positive correlation.

Other smoking variables in respect of obtaining information and having lessons about smoking (CgIn1 and LsSmk respectively) were not as highly correlated with drug-trying response variables as the 12 smoking variables previously mentioned. Moreover, regarding the variable on the number and types of people a student knows who smoke (CgEstim), it was observed that if people were other relatives and family members, it lowered the likelihood of students to try drugs, but if they were friends, it, in turn, increased such likelihood. In addition, from the variable related to getting helpful information about smoking cigarettes from people (CgPe1) and getting helpful information about smoking cigarettes from media (CgIn1), it was found that the students who obtained information from professionals and police were much less likely to try most of the drugs, and those who obtained information from interactive media such as the Internet were much less likely to try tranquillisers and heroin. From the variable in respect of the types of people the students know who smoked cigarettes (CgPp1), those students who knew other relatives, friends or all that smoked cigarettes were less likely to try drugs. However, those students who knew family member who smoked cigarettes were more likely to try drugs.

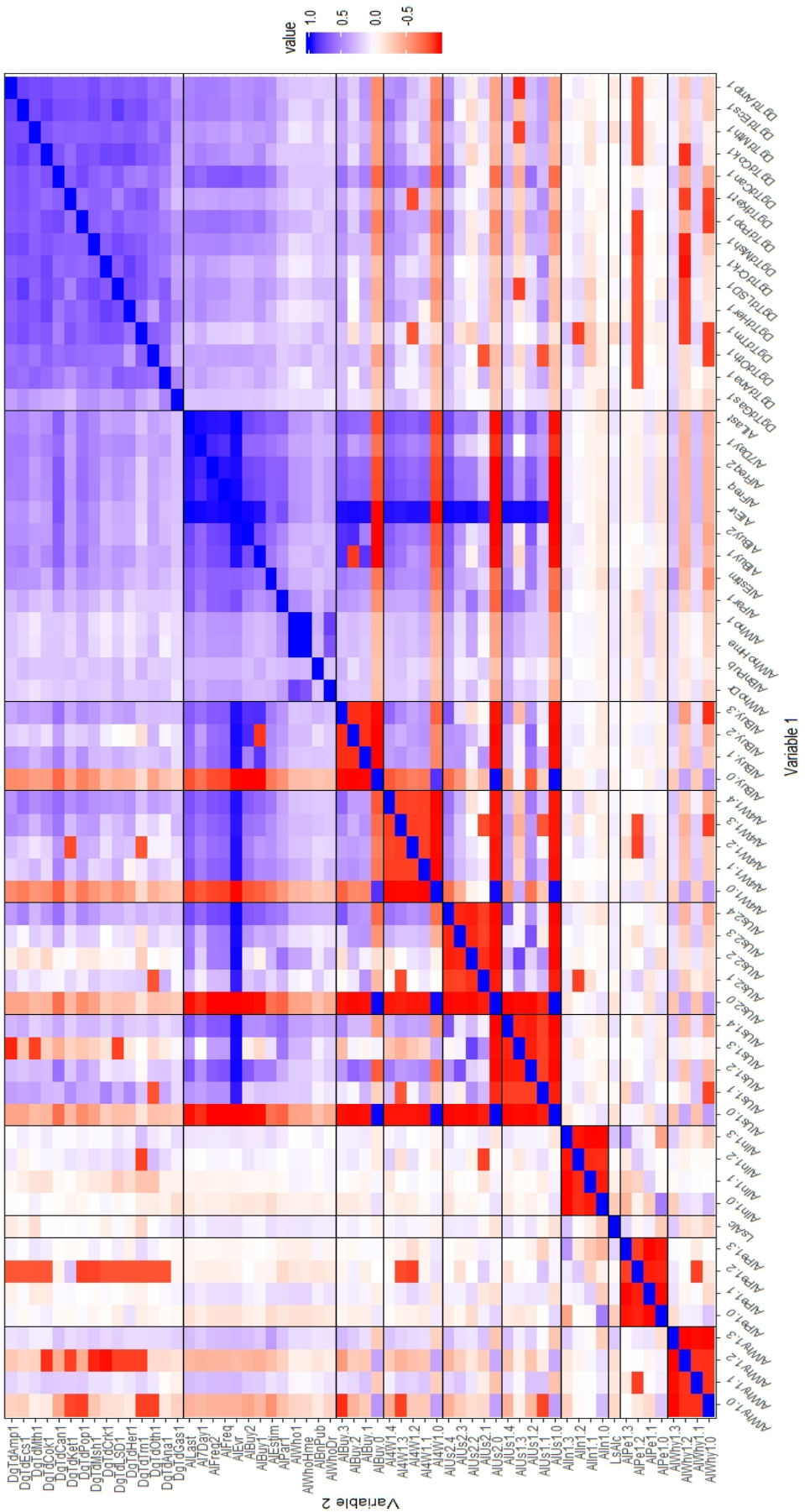

Figure 3.4: Polychoric Correlation Plot among the Drinking and Drug-trying Response Variables. Drinking variables were indicated as from AlWhy1 to AlLast, at lower rows and columns at the left, drug-trying response variables are indicated as from DgTdGas1 to DgTdAmp1, at upper rows and columns at the right. Red indicated negative correlation and blue indicated positive correlation.

Secondly, in Figure 3.4, it was observed that the 15 drug-trying response variables were strongly and positively correlated with each other and that most of the drinking variables apparently showed positive correlations with drug-trying response variables. These drinking variables were AlUs1, AlUs2, Al4W1, AlBuy, AlPar1, AlEstim, AlBuy1, AlBuy2, AlEvr, AlFreq, AlFreq2 and Al7Day1. Also, the AlLast variable was highly correlated with drug-trying response variables. These findings imply that, places of consuming alcohol, companions who drank alcohol, types of incidences a student encountered when drinking alcohol, places of purchasing alcohol, frequency of drinking alcohol, family's attitude towards alcohol consumption and the proportion of people a student knows were drinkers, are all positively associated with drug-trying among students. These drinking variables were also found highly correlated among themselves. Another observation was that from the variable AlPe1 the students who obtained information about drinking alcohol from professionals and police apparently lowered the likelihood for the students to try drugs. Compared with the smoking variables, other drinking variables in respect of obtaining information and having lessons about drinking (LsAlc, AlIn1) were not as highly correlated with drug-trying response variables as the former cluster of smoking variables. In addition, regarding the drinking variable about the reason that the students thought people of same age smoke (AlWhy1), Figure 3.4 showed that the students who thought people drank to feel better themselves were less likely to try drugs. This suggests that seeking of exuberance may be a reason for trying drugs. It was also noted that the drinking variables on a whole were positively correlated with each other. Absolute negative correlations only existed between levels of the same drinking variables.

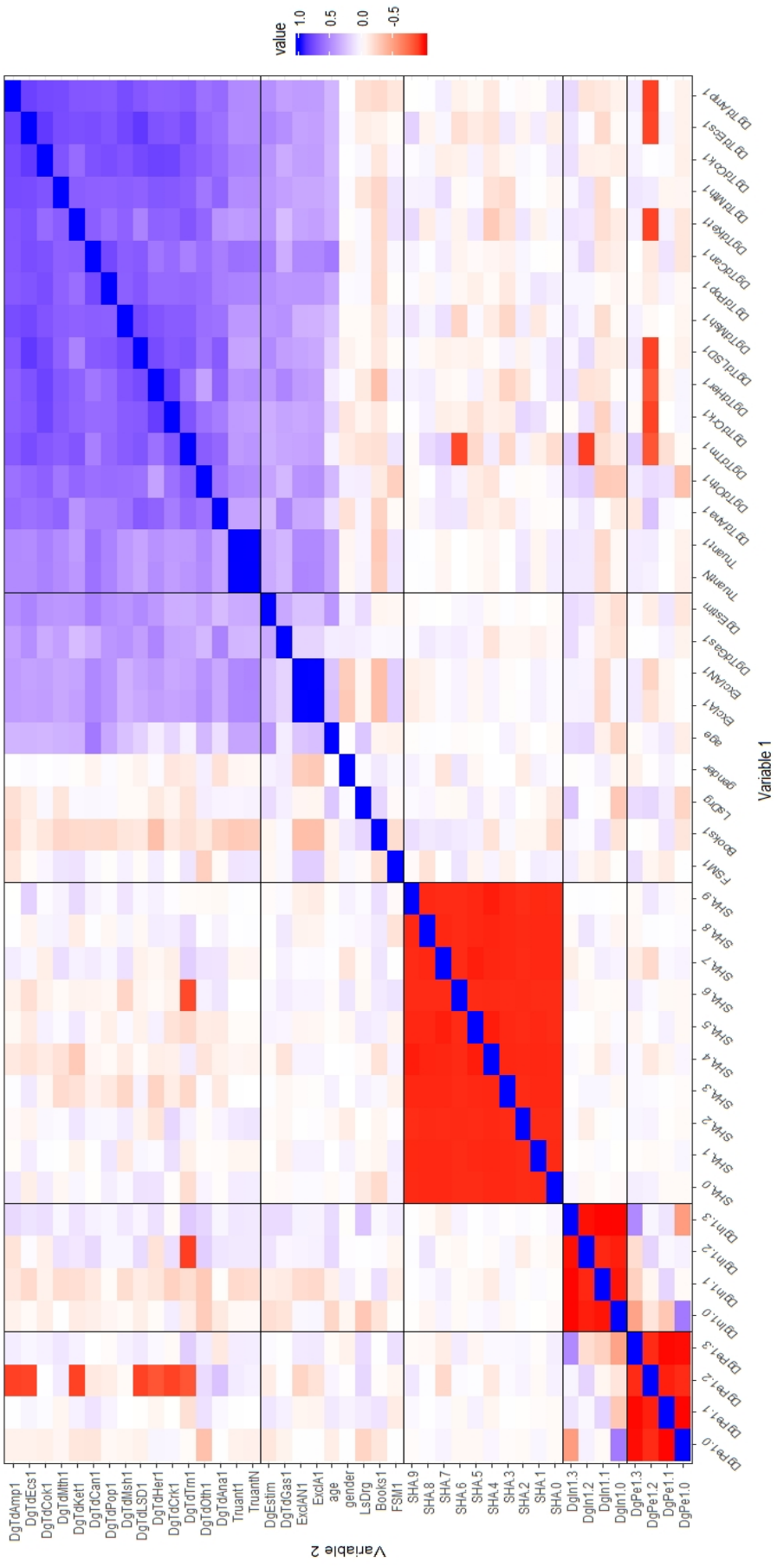

Figure 3.5: Polychoric Correlation Plot among the Drug-related Socio-demographic Variables and Drug-trying Response Variables. Drug-related Socio-demographic variables were indicated as from DgPe1 to Truant1, at lower rows and columns at the left, drug trying variables were indicated as from DgTdGas1 to DgTdAmp1, at upper rows and columns at the right. Red indicated negative correlation and blue indicated positive correlation.

Finally, in Figure 3.5, apart from the 15 drug-trying response variables were strongly and positively correlated with each other, it was observed that several drug-related socio-demographic variables showed apparent positive correlations with drug-trying response variables. Age, ExclA1, ExclAN1, DgEstim, TruantN and Truant1 were all strongly associated with drug-trying response variables, indicating that a rise in levels in age, frequency of truancy and exclusion from school, which are both measures of anti-social behaviour, increases the likelihood for a student to try drugs. In addition, students are more likely to try drugs as they get older. Another observation was that from the DgPe1 variable, the students obtained information about drugs from professionals and police were apparently less likely to try drugs. Variables in respect of drug education and information from media (LsDrg, and DgIn1 respectively), gender (Gender) and number of books (Books1) a student owns were apparently not related to drug-trying response variables.

3.3.5 Comparison with the Findings in the Year 2010 Survey Report

Compared with the findings in the Year 2010 Survey report (a summary was presented in Section 1.3.3), additional main findings of the drug-trying behaviour among young people in England by the study of the associations and relationships among drug-trying response variables and covariates (i.e. the smoking, drinking and drug-related socio-demographic variables) are summarized as follows:

(1) Similar to the finding in the Year 2010 Survey report, results of the percentage contingency tables, box plots and polychoric correlation plots consistently show the strong positive association between smoking and drug-trying behaviour of the students in England and that there are different patterns of pairwise

associations between the smoking variables and the 15 individual drugs. Results of the percentage contingency tables, box plots and polychoric correlation plots further reveal that the strong positive association between smoking and drug-trying behaviour of the students in England is highly contributed by the following smoking covariates: (1) the attitude of the students' family towards smoking (CgFam1); (2) the students' cigarette smoking status (CgStat1); (3) number of cigarettes smoked by the students in the previous week (Cg7Num); (4) frequency of purchasing cigarettes from shops by the students (CgBuyF1); (5) sources of obtaining cigarettes by the students (CgGet); (6) whether there were smokers inside the students' houses (CgWho1) as well as (7) the proportion of people a student knows who smoke (CgEstim).

(2) Similar to the smoking variable, results of the percentage contingency tables, box plots and polychoric correlation plots are consistent to the finding in the Year 2010 Survey report that there is a positive association between drinking alcohol and drug-trying behaviour of the students in England and that there are different patterns of pairwise associations between the drinking variables and the 15 individual drugs. Results of the percentage contingency tables, box plots and polychoric correlation plots further reveal that the positive association between drinking and drug-trying behaviour of the students in England is mainly contributed by the following drinking covariates: (1) the attitude of the students' family toward drinking alcohol (AlPar1); (2) usual frequency of drinking alcohol by the students (AlFreq2); (3) sources of buying alcohol by the students (AlBuy); (4) whether there were drinkers inside the students' houses (AlWho1); (5) types of incidences when the students drank alcohol (Al4W1) as well as (6) the proportion of people a student knows who drank alcohol (AlEstim).

(3) For the drug-related socio-demographic variables, results of the percentage contingency tables, box plots and polychoric correlation plots support the

findings in the Year 2010 Survey report that the drug-related socio-demographic variables, namely (a) age of the students (Age); (b) how often the students had been excluded from schools (ExclAN1) and (c) how often the students played truant (Truant1), are strongly and positively associated with drug-trying response variables. However, these three drug-related socio-demographic variables exert different patterns of pairwise associations with the 15 individual drugs. These three drug-related socio-demographic variables are particularly strongly correlated with the five drugs: (1) cannabis; (2) poppers; (3) cocaine; (4) ecstasy and (5) magic mushrooms.

(4) For the drug-trying response variables, results of the polychoric correlation plots show that the 15 drug-trying response variables are strongly and positively correlated with each other.

(5) The Year 2010 Year Survey report stated that "girls were less likely than boys to have taken drugs in the last year" (Fuller et al., 2011). According to the percentage contingency table in respect the gender variable (Gender), it reveals that the aforesaid statement is valid for seven drugs (cannabis, magic mushrooms, crack, LSD, ketamine, anabolic steroids and tranquillisers) of which the proportion percentages of male students trying them were slightly higher than female students. On the other hand, for the other eight drugs (heroin, cocaine, methadone, ecstasy, amphetamines, poppers, gas and other drugs), results of the percentage tabulate show the opposite. Similarly, the Year 2010 Survey report stated that the school-level variable (percentage of pupils eligible for the free school meals) was not significantly associated with drug use in the survey. However, the percentage contingency table in respect of whether the students have enrolled in free school meal scheme (FSM1) indicates that the students involved in the free school meal scheme are more likely to try cannabis, heroin, cocaine, magic mushrooms, methadone, ketamine, gas and tranquillisers.

The above additional main findings in the exploratory data analysis of the working data set provide hints to justify our planned effort in this research as elaborated in Section 1.4.2. To enrich the understanding of drug-trying behaviour among young people in England, development and application of advanced statistical methodologies are needed to further investigate the interactions among drug-trying response variables as well as to further study the associations among drug-trying response variables and the smoking, drinking and drug-related socio-demographic variables in the working data set.

3.4 Summary

This chapter has summarised the results of the exploratory data analysis in respect of the working data set of this research. There were 25.48% of the students who had ever smoked, and 44.83% of the students who had ever drunk alcohol. Most family members were either against or neutral towards smoking and drinking behaviour of the students. Most students knew surrounding people who either smoked or drank or took drugs, and most of them had lessons about smoking, drinking and drugs. Regarding the usage of drugs, cannabis was the most used drug, of which 9.06% of the students used it, whereas tranquillisers was the least used drug, of which only 1.85% of the students used it. A large number of the students had never tried drugs, but there were still a substantial number of the students who had tried drugs, including a few who had tried more than six drugs.

Regarding the pairwise associations between drug-trying response variables and covariates, except CgWhoSmo, gender and free school meal covariates, in general, most of the smoking, drinking and socio-demographic covariates were positively associated with drug-trying response variables. Drug-trying response

variables were also strongly and positively associated with each other.

Also, empty cells existed in some combinations of covariates and drug-trying response variables. This problem is needed to be addressed in Chapter 5 under logistic regression models.

When compared with the findings of the Year 2010 Survey report, examination of the pairwise associations and relationships in respect of drug-trying response variables and covariates (i.e. the smoking, drinking and socio-demographic variables) of the working data set by percentage contingency tables, box plots and polychoric correlation plots shed additional light to help understanding more about the drug-trying behaviour of the students. These additional findings (as summarised in Section 3.3.5) were not found in the Year 2010 Survey report. In Chapter 4, we continue our analysis by investigating the missingness of the working data set. However, before such investigation, we discuss the missing data theory applied in the working data set in Section 4.1.

Chapter 4

Missing Data Theory, Methodology and Application

4.1 Overview of Missingness

Missingness occurs for various reasons. For item non-response, reasons may include: (1) a respondent may not understand the question; (2) a respondent does not wish to answer the sensitive question; or (3) a respondent cannot figure out which option to choose in the case of multiple-choice questions. Moreover, if survey questions are deemed too tedious or too sensitive to answer, a respondent may refuse to answer (Tourangeau and Yan, 2007). Also, the internal routing system in a questionnaire may be another reason for item non-response. For unit non-response, a respondent may either have no interest or refuse to provide answers to the questionnaire or is unable to be interviewed due to language barrier and disabilities (Lavrakas, 2008).

Missing data are ubiquitous in societal and behavioural science studies (Little and Schenker, 1995), as well as in most medical, clinical and epidemiological research studies (White et al. (2009); Sterne et al. (2009); Tu and Greenwood (2012)), and are prevalent in large-scale surveys, including the "Health Survey

for England" and "Smoking, Drinking and Drug Use among Young People in England" survey series. De Leeuw et al. (2008) listed the possible causes of the increase in missing data in surveys, which include: (1) respondents are not having an answer to the question and (2) respondents' refusal to provide a response. Regarding the Year 2010 Survey employed in this research, the probable causes of the missing data included the followings: (1) the survey questionnaire contained sensitive questions which the students refused to answer and (2) a portion of the students in the survey might possess insufficient information to answer some questionnaire questions (Kyureghian et al., 2011).

The problem of missing data is a major issue in statistical analyses. Schafer and Graham (2002) stated that since most statistical analyses are not designed to deal with missing values, the occurrence of missingness hampers the statistical analysis of scientific research. If missing data are not managed properly, missingness can lead to problems of bias in the statistical estimates and a loss of efficiency (White et al. (2009); Sterne et al. (2009); Carpenter and Kenward (2013)). Despite these problems, many researchers mistreat missingness by either treating missing values as merely another category or ignore the issue of the missing data and conduct a complete case analysis instead. In order to better understand the reason for the presence of missing data, there is a need to discuss the missing data theory and mechanisms, as well as methods to properly deal with missing data.

4.2 Terminology and Models Used

Let Y be a dependent variable with missingness, where $Y = \{Y_{miss}, Y_{obs}\}$, and X be the covariates, where $X = \{X_{miss}, X_{obs}\}$. In this chapter, we examine the missingness mechanisms suggested by Rubin (1987) and investigate ways of using these mechanisms within a statistical model. When performing a statistical analysis

in the presence of missing data, the following three models are considered:

4.2.1 Substantive Model

The substantive model which concerns addressing the questions of interest, for example, finding the factors that attribute to drug-trying behaviour. The probability of the dependent variable is expressed as: $f(Y \mid X, \Theta)$, where Θ is the set of parameters of the substantive model.

4.2.2 Missingness Model

The missingness model is used to diagnose the mechanism by which the data is missing. Let $M = \{M_1, \ldots, M_N\}$ be the binary missingness indicators for Y, for respondent $1, \ldots, N$, where N represents the number of respondents. The probability of missingness in Y can be expressed as: $f(M \mid X, Y, \phi)$, where ϕ is the set of parameters of the missingness model for Y. Missing indicators for X can be defined in a similar way.

4.2.3 Imputation Model

The imputation model which formulates the methodology for imputing the data for data analysis. The probability of the covariates is expressed as: $f(X \mid \psi)$, where ψ is the set of parameters of the imputation model.

Sections 4.3 and 4.4 below provide an overview on the missingness mechanism and its implication, as well as techniques for diagnosing the missingness.

4.3 Missing Data Mechanism

Lavrakas (2008) suggested that the assumption regarding the missingness mechanism has an influential consequence for the subsequent data analysis. Accord-

ing to Rubin (2002), the missing mechanism can be defined by the assumed relationship between missingness and the values of variables in the data set. There are three classifications of missingness, namely: (1) missing completely at random (MCAR) (Rubin, 2002); (2) missing at random (MAR) (Rubin, 1976) and (3) missing not at random (MNAR) (Rubin, 1976).

MCAR is defined with an assumption that the missingness in a data set is independent of both observed and missing values in the data set. For example, a respondent flips a coin or throws a dice to decide whether he or she should respond to a question or not. MCAR is expressed by the following equation:

$$f(M \mid X, Y, \phi, \Theta, \psi) = f(M \mid \phi). \tag{4.1}$$

The existence of the MCAR mechanism can be tested through using a hypothesis test, involving logistic regression, known as Ridout's test (Ridout and Diggle, 1991), which was adopted to assess the relationship between the dropout of individuals in a clinical trial and a covariate. The Ridout's test can be applied to assess the relationship between the missingness of a covariate and observed values of another covariate. Suppose the data set X contains P covariates, i.e. $X = \{X_1, \ldots, X_P\}$, where $X_p, p = 1, \ldots, P$ represents the covariate vector for each covariate p in the data set. Suppose the missingness of a variable $p', p' = 1, \ldots, P$, $\phi = \{\phi_0, \phi_1\}$, where ϕ_0 represents the intercept of the logistic regression model, and ϕ_1 represents a parameter for covariate p in the data set. The formulation of the Ridout and Diggle (1991) test is expressed as follows:

$$\text{logit}(M_{p'}) = \phi_0 + \phi_1 X_p, p, p' = 1, \ldots, P, p \neq p'. \tag{4.2}$$

If the coefficient of ϕ_1 is determined by Wald's test (Wald, 1945) to be significant at 5 % significance level, it indicates that the missingness of Y is significantly re-

lated to X, implying that the missingness of variable Y is not MCAR. On the other hand, if the coefficient of ϕ_1 is determined by Wald's test to be not significant at 5 % significance level, we do not reject the hypothesis that the missingness of Y is not related to X. However, this does not necessarily mean the hypothesis that the missingness of variable Y is MCAR should be assumed.

Rubin (1976) defined MAR as where the missingness can be explained in terms of observed data. For example, older respondents may be more likely to refrain from reporting their income in a questionnaire. The equivalent equation of MAR is expressed as follows:

$$f(M \mid X, Y, \phi, \Theta, \psi) = f(M \mid Y_{obs}, X_{obs}, \phi), \tag{4.3}$$

where ϕ is a set of parameters causing the missingness of Y and Θ is a set of parameters in a substantive model, as stated in Section 4.2.1.

The MCAR assumption is stronger than the MAR assumption in the sense that it is assumed that missingness is independent of both observed and missing values. Though the MCAR assumption can be rejected, it is not possible to verify the MAR assumption with a single data set (White et al., 2009).

MNAR is defined as the status that the missingness is dependent on both observed values and missing values in the data set. For example, respondents who have committed at least one crime are less likely to respond to questions related to crimes in a questionnaire. The equivalent equation of MNAR is expressed as follows:

$$f(M \mid X, Y, \phi, \Theta, \psi) \neq f(M \mid X_{obs}, Y_{obs}, \phi). \tag{4.4}$$

One example of MNAR is the response to income questions in a survey, which are related to an income variable. If respondents who are earning more are less likely to answer such income questions, then the response of the income variable is subject to MNAR.

Rubin (1976) defined the term *ignorable* for the missing data mechanism with two following conditions that are needed to be fulfilled:

1. The missing data are under MCAR or MAR mechanism.
2. The set of parameters that govern the missingness model, ϕ, are distinct from the set of parameters that govern the substantive model, Θ.

4.4 Inferring the Missing Data Mechanism

According to Ibrahim et al. (2005), if the missingness of Y is ignorable, then the missingness model term $f(M \mid X, Y, \Theta, \phi, \psi)$ can be simplified into $f(M \mid X, Y, \phi)$, since this term represents the probability of the missingness model, involving parameters which predict the missingness of the data set. No parameters that do not cause missingness are included in this term. The likelihood term $f(Y \mid X, \Theta, \phi, \psi)$ can be simplified into $f(Y \mid X, \Theta)$, since this term represents the likelihood of the substantive model, involving parameters which predict the response variable and parameters that are related to the response variable. The covariate term $f(X \mid \Theta, \phi, \psi)$ can be simplified into $f(X \mid \psi)$, since the covariate term only depends on parameters that are related to covariates.

The joint likelihood of the data set, the covariate terms and the missingness are represented by the selection model as described by the following equation:

$$f(X,Y,M \mid \Theta,\phi,\psi) \propto f(M \mid X,Y,\phi) f(Y \mid X,\Theta) f(X \mid \psi). \tag{4.5}$$

(Ibrahim et al., 2005)

An alternative factorisation of the joint likelihood expressed in Equation 4.5 is a pattern-mixture model (Little, 2008). The pattern-mixture model is expressed as follows:

$$f(X,Y,M \mid \Theta,\phi,\psi) \propto f(Y \mid X,M,\Theta) f(X \mid M,\psi) f(M \mid \phi). \tag{4.6}$$

(Little, 2008)

Here, the pattern-mixture model specifies the marginal distribution of the missingness indicators, as well as the conditional distribution of the response variables based on missingness indicators. More details of the selection model and the pattern-mixture model can be referred to Glynn et al. (1986) and Little (1993). The pattern-mixture model is only feasible when there are a small number of missingness patterns. Under a large number of missingness patterns, the application of the pattern-mixture model is difficult.

There is no test to determine whether the missingness is MAR. However, Allison (2001) stated that it is unlikely for this condition to be "violated in real world situations", including the "Smoking, Drinking and Drug Use among Young People in England" study series, in which as many responses were collected as possible in a strictly confidential way. As a result, treating the MAR missingness as ignorable is common. Buuren (2012) stated that for "practical purposes", the "missing data model" can be considered "ignorable if MAR holds", in the belief that the observed data "are sufficient to correct for the effects" of all "missing data" in a data set. Data sets with MAR missingness should be treated with more sophisticated imputation methods. In Section 4.5 below, we discuss the

methods and procedures for handling missing data that are MAR.

4.5 Handling Missing Data

Various ways of handling missing data that are MAR have been widely used by researchers through imputation to assess model uncertainty. In this section, details of the imputation model specifications and inferential frameworks used in this work are described. The two imputation model specifications mentioned in this section are: (1) joint models, which is described in Section 4.5.2, and (2) fully conditional specification, which is described in Section 4.5.3. The full conditional specification is used in the `mice` package for multiple imputation by chained equations.

Before describing these two imputation model specifications, single imputation is briefly introduced in Section 4.5.1.

4.5.1 Single Imputation

Previously, single imputation models were often used to deal with missing data, including listwise deletion (Kelejian, 1969), pairwise deletion (Schafer and Olsen, 1998) and predictive mean matching (Little, 1988). However, in a single imputation, as the name suggests, the missing data are only imputed once. The uncertainty due to the need for imputation is therefore neglected in a single imputation. Moreover, in a single imputation, there is no method of differentiating between the imputed data and the non-imputed data in the subsequent analysis. This problem is addressed by imputing stochastically the missing data more than once, ideally more than five times. The method of imputing the missing data repeatedly is called multiple imputation, which is introduced in the next two sections.

4.5.2 Joint Models

Multiple imputation by joint models assumes that the data can be described by a multivariate distribution, depending on the type of data. The limitation of this approach is that the data rarely conform to a predefined joint distribution, and transformation is often required to address this discrepancy (Buuren (2007); Buuren (2012)). This implies a substantial amount of effort to identify suitable joint distributions and transformations for all variables with missing values in a large data set. One alternative approach is to implement multiple imputation through fully conditional specification, which is described in Section 4.5.3 below.

4.5.3 Full Conditional Specification (FCS)

In contrast to specifying a joint model for the data, multiple imputation by fully conditional specification (FCS) (Buuren, 2012) adopts pre-specified conditional distributions for each variable in a data set. Each of these variables is then imputed sequentially. When specifying the full conditional distributions for each variable, the data model must be defined, according to the type of data, and is taken to be conditional upon the other variables in the data set.

One issue with FCS is compatibility (Buuren, 2012). Another issue with FCS is convergence. Buuren and Groothuis-Oudshoorn (2011) suggested that convergence is achieved when the following two criteria are achieved: (1) different chains (i.e. posterior chains) of imputation are freely and stably mingled with each other and (2) the 'between-imputation' variance between different chains is not larger than the average within-imputation variance.

Both the fully Bayesian framework and the multiple imputation by chained equations are sampled from full conditional distributions. The fully Bayesian framework is a one-stage modelling approach, since the imputation model and the

substantive model are integrated, whereas the multiple imputation by chained equations is a two-stage modelling approach, since the imputation model and the substantive model are separately conducted in two stages. The fully Bayesian framework is described in Section 4.5.3.1, whereas the multiple imputation by chained equations (Stage 1) and response modelling (Stage 2) are described in Section 4.5.3.2 and 4.5.3.3 respectively.

4.5.3.1 Fully Bayesian Framework

Both Spiegelhalter (2003) and Lunn et al. (2006) discussed the fully Bayesian framework. The basic idea of a fully Bayesian framework for missing data involves specifying priors for all parameters concerned, along with specifying distributions for the missing data. These missing data are sampled from their conditional posterior distribution through a Gibbs sampler (Geman and Geman (1984); Spiegelhalter (2003)).

A fully Bayesian framework is an approach that not only imputes the missing data, but also models the uncertainty of the imputation. It is a one-stage method, since missing data and parameters are imputed within the same statistical model. In a fully Bayesian framework, each parameter is given a prior and all the parameters are initialised by drawing initial values from their corresponding priors. These parameters are then updated from their corresponding posterior functions, while also sampling the missing data from the likelihood of the imputation model, until convergence. In the case that there is missing data in the covariates, then a prior must be placed on these covariates.

The conditional prior for the covariate X, which is based on the parameter of covariates of the imputation model ψ, is denoted by $p(X \mid \psi)$. After observing the response data, Y, the prior is updated from a posterior. The probability density function for the response data Y, depending on the parameter Θ, is

called likelihood function ($p(Y \mid X, \Theta)$), which can be used to impute the missing data. When there is missingness in the covariates, a prior distribution must be specified for the covariates as a part of the imputation model. In this case, the prior for the covariates will become $p(\psi)$.

In the fully Bayesian framework, the joint probability model for Y, X, M, Θ, ψ, ϕ is as follows:

$$p(Y, X, M, \Theta, \phi, \psi) \propto p(Y \mid X, \Theta) p(X \mid \psi) p(M \mid Y, X, \phi) p(\Theta) p(\psi) p(\phi), \tag{4.7}$$

(Ibrahim et al. (2005);Best and Mason (2012))

where $p(Y \mid X, \Theta)$ is the likelihood of the substantive model, $p(X \mid \psi)$ represents the likelihood of the imputation model and $p(M \mid Y, X, \phi)$ represents the missingness model, $p(\Theta)$, $p(\phi)$ and $p(\psi)$ are priors for Θ, ϕ and ψ, respectively.

In this section, we focus on the case when the assumption of ignorability can be made. The non-ignorable case has been explained in Ibrahim et al. (2005). In other words, both ϕ and M terms can be ignored. The subsequent joint probability model for Y, X, Θ, ψ is therefore represented as follows:

$$p(Y, X, \Theta, \psi) \propto p(Y \mid X, \Theta) p(X \mid \psi) p(\Theta) p(\psi). \tag{4.8}$$

A fully Bayesian framework utilises various samplers, including the Gibbs Sampler (Geman and Geman, 1984) and Slice Sampler (Ntzoufras, 2009) to update the parameters and missing values. This method begins with drawing the initial values for ψ and Θ parameters, as well as missing values of X and Y, by specifying the distribution terms of the missing values of X and the priors of ψ and Θ. The missing values X_{miss}, Y_{miss}, ψ and Θ are all sampled from their respective conditional terms at each iteration t, $t = 1, \ldots, T$. The algorithm of

sampling by the Gibbs Sampler is provided as Algorithm 4.1 below:

Algorithm 4.1 Fully Bayesian Framework

1: Specify the imputation model X, $p(X \mid \psi)$ and the likelihood function, $p(Y \mid X, \Theta)$.
2: Specify the prior distributions for parameters of the substantive model and the imputation model, Θ and ψ respectively.
3: Initialise ψ, Θ, X_{miss} and Y_{miss} as $\psi^{(0)}$, $\Theta^{(0)}$, $X_{miss}^{(0)}$ and $Y_{miss}^{(0)}$.
4: **for** $t = 1, \ldots, T$ **do**
5: Sample ψ, Θ, X and Y from the joint distribution function, $p(Y, X, \Theta, \psi)$, to the following conditional posteriors:
a: $\Theta^{(t)}$ from $p(Y^{(t-1)} \mid X^{(t-1)}, \Theta^{(t)})p(\Theta^{(t)})$ - to sample from this distribution, propose $q(\Theta^{(t)})$ and accept according to Metropolis-Hastings ratio;
b: $\psi^{(t)}$ from $p(X^{(t-1)} \mid \psi^{(t)})p(\psi^{(t)})$ - to sample from this distribution, propose $q(\psi^{(t)})$ and accept according to Metropolis-Hastings ratio;
c: $Y_{miss}^{(t)}$ from $p(Y^{(t)} \mid X^{(t-1)}, \Theta^{(t)})$;
d: $X_{miss}^{(t)}$ from $p(X^{(t)} \mid \psi^{(t)})$.
6: **end for**

Steps 5a to 5d of Algorithm 4.1 are repeated for all Θ and ψ terms, all Y_{miss} missing responses and all X_{miss} missing covariates, until convergence is obtained for all these terms altogether. The methods for convergence check (i.e. trace plots containing mean estimates and standard errors) are demonstrated in Section 4.7.

Calculating the conditional terms in Steps 5a to 5d of Algorithm 4.1 can be a technically demanding and complex task. `WinBUGS` and `OpenBUGS` programmes (Lunn et al. (2000); Spiegelhalter (2003); Spiegelhalter (2009)) make this computing task much simpler for the user. In `WinBUGS`, only the priors, $p(\Theta)$ and $p(\psi)$, and the likelihood, $p(X \mid \psi)$ and $p(Y \mid X, \Theta)$, are needed to be specified, and the conditional terms are sampled from the specified priors and likelihoods automatically.

In the `OpenBUGS` program (Spiegelhalter, 2009), the Gibbs sampling (Geman and Geman, 1984) is the main sampler of the fully Bayesian imputation, in which the component imputation involves various methods, including rejection sampling

and Slice sampler. Further details about Slice sampler and rejection sampling can be found in Ntzoufras (2009). In this research, `OpenBUGS` program was used because the environment was easy to use for writing Bayesian imputation programs (Murphy, 2007).

4.5.3.2 Multiple Imputation by Chained Equations - First Stage

Multiple imputation by chained equations creates many copies of the fully observed data with varying imputed values for missing data. It is a two-stage imputation model, as the missing data are imputed by an imputation model, and statistical inferences are implemented on imputed data sets with a substantive model. Section 4.5.3.3 discusses how the models on the imputed data sets are combined. In this section, we show how the multiple imputation by chained equations can be used to carry out multiple imputation.

The multiple imputation by chained equations (MICE)(Buuren, 2012) is an implementation of fully conditional specification (FCS), which imputes missing data on a "variable-by-variable" basis, for T iterations, P number of variables in the data set and W imputed data sets. The MICE imputation deals with parameters for substantive model at iteration t of the Gibbs sampler, denoted as $\Theta_{w,p}^{(t)}$, and parameters for imputed model denoted as $\psi_{w,p}^{(t)}$ for $t = 1,\ldots,T, w = 1,\ldots,W$. This imputation deals with each parameter sequentially, which are $\Theta_{w,p}, \psi_{w,p}, p = 1,\ldots,P, w = 1,\ldots,W$. The missing data $X_{w,p,miss}$ is sampled from $X_{w,p,miss}^{(t)} \sim P(X_{w,p,miss}^{(t)} \mid \psi_{w,p})$ for $t = 1,\ldots,T, w = 1,\ldots,W, p = 1,\ldots,P$. With implementation of FCS and Gibbs Sampling (Geman and Geman (1984);Buuren (2012)), the MICE sampling procedure can be described with the following equations and the steps of MICE algorithm for imputation of multivariate missing data are listed as Algorithm 4.2 of the first stage of modelling (Buuren, 2012).

Buuren and Groothuis-Oudshoorn (2011) developed the `mice` package in `R` pro-

Algorithm 4.2 First Stage of MICE Modelling Algorithm

for Imputation $w = 1, \ldots, W$ **do**
    Specify the imputation model $X_{w,miss}$, $p(X_{w,miss} \mid \psi_w)$ and the likelihood function, $p(Y_w \mid X_w, \Theta)$
    Initialise $Y_{w,miss}$ as $Y^{(0)}_{w,miss}$.
    for All parameters $p = 1, \ldots, P$ **do**
        Initialise $X_{w,p,miss}$ as $X^{(0)}_{w,p,miss}$.
        Initialise $\Theta_{w,p}$ as $\Theta^{(0)}_{w,p}$.
        Initialise $\psi_{w,p}$ as $\psi^{(0)}_{w,p}$.
    end for
end for
for Imputation $w = 1, \ldots, W$ **do**
    for $t = 1, \ldots, T$ **do**
        $Y^{(t)}_{w,miss} \sim P(Y^{(t)}_{w,miss} \mid \ldots)$;
        for $p = 1, \ldots, P$ **do**
            $X^{(t)}_{w,p,miss} \sim P(X^{(t-1)}_{w,p,miss} \mid \psi_{w,p})$;
            $\Theta^{(t)}_{w,p} \sim P(\Theta^{(t)}_{w,p} \mid \ldots)$;
            $\psi^{(t)}_{w,p} \sim P(\psi^{(t)}_{w,p} \mid \ldots)$;
        end for
    end for
end for
for $w = 1, \ldots, W$ **do**
    Upon convergence, obtain the last imputed data set: $Y^{(T)}_{w,miss}$ as the w^{th} imputed data set, Y_w, and proceed to Second stage MICE algorithm.
end for

gram (R Version 3.3.0). This program calculates posterior probabilities based on generalized linear models. For binary data response, we specify a logistic regression model; for nominal variable with more than two levels, we specify a polytomous regression model, and for continuous data, we specify a linear regression with prediction method. Details of this program can be referred to the R program manual (R Development Core Team, 2008) about `mice` package (Buuren and Groothuis-Oudshoorn, 2011).

4.5.3.3 Multiple Imputation by Chained Equations - Second Stage

The second stage of modelling fits the substantive models on imputed data sets, followed by combining estimates of the substantive models through Rubin's rule

(Rubin, 1987). Through the substantive model for each w^{th} of the W imputed data sets, $w = 1, \ldots, W$, a set of parameter estimates, $\hat{\Theta}_w$ and its covariance matrix, $\hat{V}_w = Var(\hat{\Theta}_w)$, are obtained for variable Θ_p. Estimates and variances from W imputed data sets are combined by Rubin's rule (Rubin, 1987). Rubin's rule is applicable to imputed data sets under MAR, ignorability and normality assumptions (Allison, 2003).

Suppose W sets of estimates are obtained from analysis of W imputed data sets for the variable Θ_p, denoted as the estimate vector $\hat{\Theta}_w, w = 1, 2, \ldots, W$, the combined mean estimate vector for W imputed data sets, $\bar{\Theta}_W$, is calculated by Equation 4.9:

$$\bar{\Theta}_W = \frac{1}{W} \sum_{w=1}^{W} \hat{\Theta}_w. \tag{4.9}$$

Regarding the calculation of total combined covariance matrix, firstly we explain the calculation of the combined within-imputation covariance matrix, $\bar{V}_W$, and then the combined between-imputation covariance matrix, $\bar{B}_W$. Finally, we explain the calculation of the total combined covariance matrix, $\bar{T}_W$. At this stage, we utilise W number of covariance matrices, $\hat{V}_1, \ldots, \hat{V}_W$, that are associated with estimate vectors $\hat{\Theta}_1, \ldots, \hat{\Theta}_W$ respectively.

The combined within-imputation covariance matrix, $\bar{V}_W$, is calculated as in Equation 4.10, which is basically the mean of variances for $\hat{\Theta}_w$, $w = 1, \ldots, W$, from W imputed data sets:

$$\bar{V}_W = \frac{1}{W} \sum_{w=1}^{W} \hat{V}_w. \tag{4.10}$$

The combined between-imputation covariance matrix, $\bar{B}_W$, is calculated as in Equation 4.11:

$$\bar{B}_W = \frac{1}{W-1} \sum_{w=1}^{W} (\hat{\Theta}_w - \bar{\Theta}_W)(\hat{\Theta}_w - \bar{\Theta}_W)^T. \tag{4.11}$$

Hence, the total combined covariance matrix, $\bar{T}_W$, is calculated as in Equation 4.12:

$$\bar{T}_W = \bar{V}_W + \left(1 + \frac{1}{W}\right)\bar{B}_W, . \tag{4.12}$$

(Rubin, 1987)

and Rubin's combination rule hence provides an unbiased estimate of the total combined covariance matrix.

Wald's test is used to determine whether a variable is significantly related to responses. For combined estimates, Wald's test statistics are adopted for testing a certain variable, Θ_p, which contains k' components to be tested. Suppose $\bar{\Theta}_W$ is the mean estimate vector for Θ_p over W imputed data sets in Rubin's rule equations, $\bar{\Theta}_0$ is the vector of null values for testing Θ_p, $\bar{V}_W$ and $\bar{B}_W$ are combined within-imputation covariance matrix and combined between-imputation covariance matrix respectively in Rubin's rule equations, the Wald's test statistic, $\omega(\bar{\Theta}_W)$, is calculated as follows:

$$\omega(\bar{\Theta}_W) = \frac{(\bar{\Theta}_W - \bar{\Theta}_0)^T \bar{V}_W^{-1} (\bar{\Theta}_W - \bar{\Theta}_0)}{(1+r)k'}, \tag{4.13}$$

where k' is the number of components being tested, $r = (1+1/W)\text{trace}(\bar{B}_W \bar{V}_W^{-1})/k'$. The p-value by F distribution is then evaluated, and the corresponding p-value is stated as follows:

$$P[F_{k',l} > \omega(\bar{\Theta}_W)], \tag{4.14}$$

where $F_{k',l}$ is a random variable of F distribution with k' and l degrees of freedom. For $k'(W-1) > 4$, l is defined as follows:

$$l = 4 + [k'(W-1) - 4]\left(1 + \frac{z}{r}\right)^2, z = \left\{1 - \frac{2}{k'(W-1)}\right\}. \tag{4.15}$$

Alternatively, $l = (W-1)(k'+1)(1+1/r)^2/2$ if $k'(W-1) \leq 4$.

(Li et al., 1991)

The Wald's test will also be applied in the backward elimination process with Rubin's rule in logistic regression model, log-linear analysis model, item response theory model and latent class analysis model.

The algorithm of the second stage of the MICE imputation is described as Algorithm 4.3.

Algorithm 4.3 Second Stage of MICE Variable Selection Algorithm (Combining Estimates)

1: **while** An insignificant covariate exists in a substantial backward elimination model **do**
2: **for** $w = 1, \ldots, W$ **do**
3: Fit the substantial model with every Y_w, and generate $\hat{\Theta}_w$, $\hat{V}_W$ as results.
4: **end for**
5: Calculate $\bar{\Theta}_W = \frac{1}{W}\sum_{w=1}^{W}\hat{\Theta}_w$.
6: Calculate $\bar{V}_W = \frac{1}{W}\sum_{w=1}^{W}\hat{V}_w$.
7: Obtain the between-imputation variance $\bar{B}_W$ and total variance $\bar{T}_W$ by Rubin's Rule.
8: Perform Wald's test with $\hat{\Theta}_w$, $\bar{\Theta}_W$, $\bar{V}_W$, $\bar{B}_W$ and $\bar{T}_W$ on each covariate and determine which insignificant one (at 5% significance level) to discard by the highest p-value.
9: Discard the insignificant covariate with the highest p-value.
10: **end while**

To determine which covariates to be included in the imputation model, polychoric correlation plots among complete cases are adopted before imputation of missing values. Generally, covariates that yield correlation value with any other covariates of 0.3 or more are included in the imputation model. Referring to Figures 3.3 to 3.5, most covariates yielded high correlation values with at least one other covariates. In addition, percentage tables and box plots in Chapter 3 reflected that most variables were related to drug response variables. As such, all covariates were included in the imputation model of MICE imputation.

4.6 Application: Building an Imputation Model

This section discusses how the fully Bayesian framework and multiple imputation methods in Section 4.5.3 can be applied to our data set of the Year 2010 Survey. We adopted two types of data set: (1) a type of data set with covariates, excluding nested variables and derived variables and (2) another type of data set with the 15 drug-trying response variables only. Sections 4.6.1 and 4.6.2 below describe how we applied fully Bayesian framework and multiple imputation by chained equations to these two types of data set.

4.6.1 Fully Bayesian Framework

To impute missing data by fully Bayesian framework for the data set with the 15 drug-trying response variables only, we used `OpenBUGS` program. Further details about the `OpenBUGS` program code can be found in Ntzoufras (2009). We specified a statistical model with parameters and equations for missing responses. We linked these parameters with observed covariates and we specified priors for these parameters. We loaded two Markov chains and compiled the data set and the statistical model. After specifying the initial values for the parameters, we updated model parameters and missing data for 17,000 cycles with 1,000 cycles of burning-in, providing 16,000 usable cycles for statistical inference. We also diagnosed the trace plots of the convergence of both Markov Chains.

The fully Bayesian Framework was applied to item response theory in Chapter 6. Details of the fully Bayesian Framework applied in item response theory can be referred to Sections 6.2.2 and 6.3.

4.6.2 Multiple Imputation by Chained Equations

In the multiple imputation by chained equations, we used `mice` package in `R` program (Buuren and Groothuis-Oudshoorn, 2011) to facilitate the multiple im-

putation by chained equations on the two types of data set. Here, two MICE imputation schemes were involved, namely scheme 1: MICE imputation scheme based on 15 drug-trying response variables only and scheme 2: MICE imputation scheme based on full data frame. We produced $W = 10$ imputed data sets through 200 imputation cycles. For binary data, we adopted logistic regression method (`logreg`); for categorical variables that contained more than two levels, we adopted multinominal (polynominal) logit regression model (`polyreg`); for continuous variables, we adopted normal linear regression model (`norm`). All these methods were under Bayesian method according to Rubin (1987) and Brand (1999).

For continuous variables with lower limits, upper limits, or both, we transformed them to approximate normality before imputation by the following methods. Suppose a continuous variable Y_P has a lower limit of zero, and we wished to transform Y_P into Y'_P for imputation, then for each value of Y_P corresponding to respondent i, $Y_{i,P}, i = 1, \ldots, N$, we adopted a transformation function $f : (0, \infty) \to \mathbb{R}$, to transform each $Y_{i,P}$ to $Y'_{i,P}$. The transformation function for each $Y_{i,P}$ was defined as below:

$$f(Y_{i,P}) = Y'_{i,P} = \log(Y_{i,P}). \tag{4.16}$$

For any $Y_{i,P} = 0$, we added a small number, i.e. 1×10^{-6}, onto $Y_{i,P}$ before applying the transformation function. After imputation, we used the inverse function $f^{-1}(Y'_{i,P})$ to transform $Y'_{i,P}$ back to $Y_{i,P}$. Values of $Y_{i,P}$ between 0 and 1×10^{-6} were treated as 0, and values of $Y_{i,P}$ between $u - 1 \times 10^{-6}$ and u were treated as u.

The above log-transformation method was implemented on any count data, as well as any variable that span across the range $[0, \infty)$ for mapping and transforming such data to $(-\infty, \infty)$ (`R` domain). The log-transformation is not

a variance-stabilising function, whereas the square root transformation is the variance-stabilising function. However, the log-transformation was chosen in this study over the square root transformation based on the following reasons: (1) square root transformation only maps variables that span monotonically across the range $[0, \infty)$ to $[0, \infty)$, given that the orders of the values are maintained; (2) it was possible that square-rooted value can be either negative value or positive value, such that values can be mapped from $[0, \infty)$ to $(-\infty, \infty)$; but in this case, this mapping is no longer monotone. However, log-transformation is monotone whlist mapping values from $[0, \infty)$ to $(-\infty, \infty)$. In other words, orders of values can be maintained during mapping and (3) we use `MICE` package in `R` programme for imputing missing data by multiple imputation by chained equations by the time of data analysis, since `MICE` package is the only `R` package that offered multiple imputation by chained equations. However, `MICE` did not offer Poisson or Negative Binomial regression option for count variables, nor Gamma regression option for variables that span across the range $[0, \infty)$.

As mentioned above that log-transformation is not a variance-stabilising function, there might be a risk of heteroscedasticity in regression analysis, compromising likelihood estimation of variable standard errors. However, as there were only three variables that used this log-transformation method for imputing missing values in MICE transformation scheme in this study, and the regression analysis that used iterative weighted least square, the same method used in the regression analysis, is robust against heteroscedasticity (Mak, 1992), log-transformation might not be a serious problem in the regression analysis.

For example, for the variable recording the number of cigarettes the students have smoked during a week prior to the survey (Cg7Num), the values before imputation and after imputation by Equation 4.16 are listed in the following table:

Table 4.6.1: Values of Cg7Num Variable during Imputation

Before imputation			After imputation		
$Y_{i,P}$	augmented $Y_{i,P}$	$Y'_{i,P}$	$Y'_{i,P}$	augmented $Y_{i,P}$	$Y_{i,P}$
0	0.000001	-13.81551	-13.81551	0.000001	0
0.5	0.5	-0.693147	-0.693147	0.5	0.5
1	1	0	0	1	1
5	5	1.609438	1.609438	5	5
10	10	2.302585	2.302585	10	10
missing	missing	missing	-20	0.000000	0
missing	missing	missing	0.5	1.648721	1.648721

We ordered the variables in both data sets, in ascending order, according to the percentage of missingness (from the smallest percentage of missingness to the largest percentage of missingness). This method was implemented since the multiple imputation by chained equations was implemented on each covariate according to its order, and arranging covariates with the fewest missing data to be imputed in higher priority led to more observed data available for imputation during imputation process, thus improving the prediction of missing values. After imputation, we checked the mean and standard deviation plots of all variables involved in the imputation to diagnose if all these variables were converged.

4.7 Application to Working Data Set

In this section, we begin with the exploration of missing data in the working data set in Section 4.7.1. Afterwards, we discuss how Bayesian and MICE imputation schemes described in Sections 4.5 and 4.6 are applied to the working data set, in Sections 4.7.2 and 4.7.3 respectively.

4.7.1 Exploration of Missing Data in Working Data Set

The exploratory analysis on missingness was conducted on the working data set, for the purpose to understand the relationship between missingness of one variable and other variables. We employed the missing data theories and mechanisms, which we have described in the previous sections of this chapter, to analyse our working data set in the aspects of exploration of missing data, construction of a model for imputation and validating imputation diagnostics. In this section, we discuss the following two types of exploratory analyses on missingness.

(1) Frequency and Percentage of Missingness in the working data set (in Section 4.7.1.1).
(2) Missingness Plots for Working Data Set (in Section 4.7.1.2).

4.7.1.1 Frequency and Percentage of Missingness

In the working data set, across all 68 variables, altogether there were 3,855 complete cases out of 7,296 available cases, accounting for 52.84 % of all cases. Among all variables in the working data set, the maximum percentage of missingness for one variable, i.e. LsDrg, was 16.98 %.

Table 4.7.1 provides information about the missingness of each of the 68 variables, sorted by the percentage of missingness in descending order. The corresponding missingness proportion bar plot for each variable is displayed in Figure 4.1.

Table 4.7.1: Frequency and Proportion of Missingness for Each Variable in the Working Data Set

Variable	Frequency (Prop.)	Variable	Frequency (Prop.)
LsDrg	1239 (16.98%)	FSM1	237 (3.25%)
LsAlc	1179 (16.16%)	Truant1	236 (3.23%)
LsSmk	1158 (15.87%)	ExclAN1	195 (2.67%)
AlIn1	546 (7.48%)	ExclA1	187 (2.56%)
AlWho1	541 (7.42%)	DgEstim	184 (2.52%)
DgPe1	539 (7.39%)	DgTdAmp1	173 (2.37%)
DgIn1	539 (7.39%)	AlEstim	173 (2.37%)
AlWhoDr	525 (7.20%)	CgNow	169 (2.32%)
CgPp1	525 (7.20%)	DgTdMth1	159 (2.18%)
CgIn1	505 (6.92%)	DgTdEcs1	158 (2.17%)
AlPe1	496 (6.80%)	DgTdMsh1	156 (2.14%)
CgPe1	457 (6.26%)	DgTdHer1	156 (2.14%)
CgWho1	428 (5.87%)	DgTdPop1	153 (2.10%)
Al4W1	419 (5.74%)	DgTdCan1	150 (2.06%)
CgWhoSmo	416 (5.70%)	DgTdCok1	149 (2.04%)
AlWhoHme	414 (5.67%)	DgTdAna1	148 (2.03%)
AlPar1	386 (5.29%)	DgTdCrk1	146 (2.00%)
CgWhoHme	357 (4.89%)	DgTdOth1	143 (1.96%)
AlWhy1	323 (4.43%)	DgTdLSD1	141 (1.93%)
Cg7Num	322 (4.41%)	DgTdGas1	137 (1.88%)
AlBuy	317 (4.34%)	DgTdTrn1	135 (1.85%)
AlBuy2	317 (4.34%)	DgTdKet1	134 (1.84%)
AlBuy1	317 (4.34%)	AlEvr	92 (1.26%)
CgBuyF1	309 (4.24%)	CgPk1	82 (1.12%)
AlUs2	290 (3.97%)	CgStat1	79 (1.08%)
Books1	285 (3.91%)	CgGet	77 (1.06%)
Al7Day1	285 (3.91%)	CgGet3	77 (1.06%)
AlUs1	282 (3.87%)	CgGet2	77 (1.06%)
AlLast	279 (3.82%)	CgGet1	77 (1.06%)
AlBnPub	278 (3.81%)	CgStat	42 (0.58%)
CgEstim	274 (3.76%)	gender	0 (0%)
TruantN	270 (3.70%)	Age	0 (0%)
AlFreq2	268 (3.67%)	SHA	0 (0%)
AlFreq	264 (3.62%)		
CgFam1	254 (3.48%)		

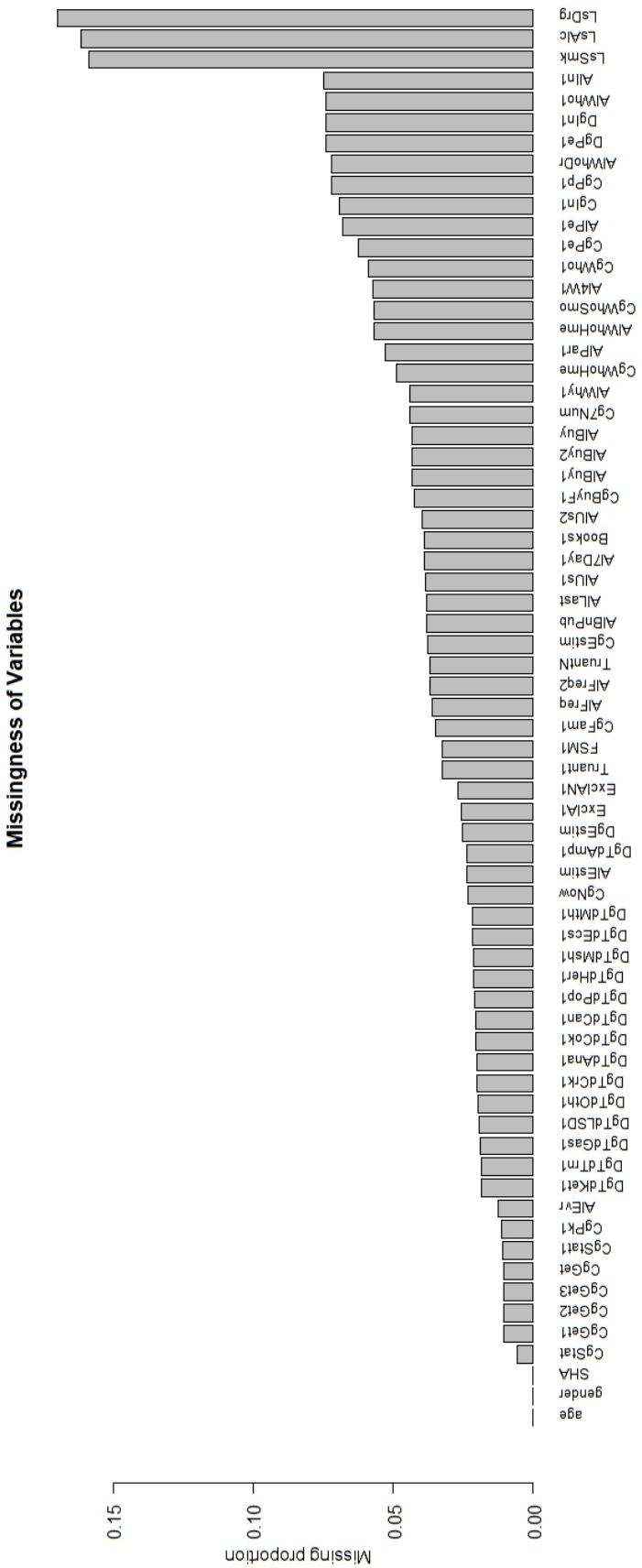

Figure 4.1: Missingness Proprtion Bar Plot for each Variable in the Working Data Set.

From Table 4.7.1, the variables representing whether a student had lessons about drug taking, drinking alcohol and smoking yielded relatively higher proportions of missingness, from 15.87 % to 16.98 %. Such high proportions existed due to the additional "don't know" option in their corresponding questions, and those "don't know" responses were treated as missing, leading to an increase in missingness. In addition, obtaining information about smoking, drinking alcohol and drug use typically yielded relatively higher proportions of missingness. From Table 4.7.1, 37 variables yielded missingness of over 3 %, indicating that missingness in the working data set was substantial, though not severe. In addition, three variables were completely observed, indicating that there were no missing values in these three variables. In a nutshell, we concluded that the missingness of every variable in the working data set was not huge, though substantial.

When we investigated the missingness of each individual student, i.e. number of missing values for each student, we examined the frequency of missing values for each student, along with missing pattern of each student. The frequency of missing value for each student is presented in Table 4.7.2, and the corresponding histogram plot is presented in Figure 4.2.

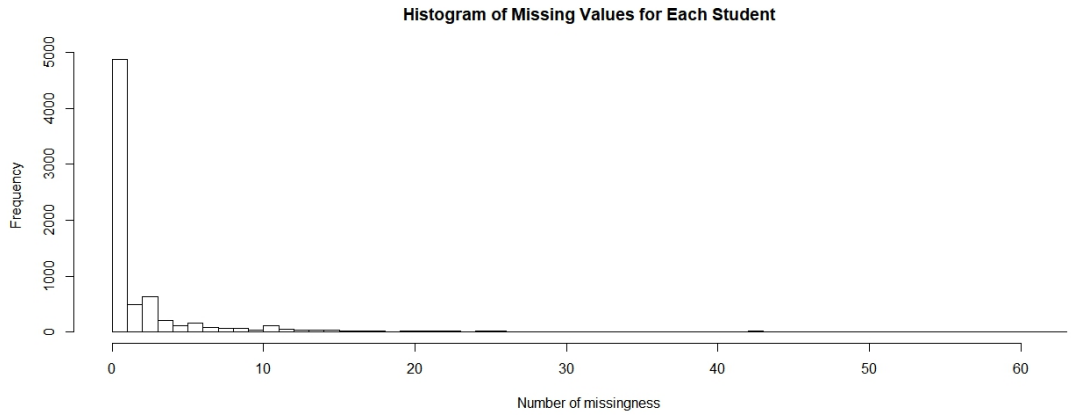

Figure 4.2: Histogram of Number of Missingness for Each Student in the Year 2010 Survey

Table 4.7.2: Table of Frequency and Proportion of Missing Values for Each Student in the Year 2010 Survey

Missing	Frequency	Missing	Frequency	Missing	Frequency
0	3855 (52.84%)	20	14 (0.19%)	40	1 (0.01%)
1	1026 (14.06%)	21	18 (0.25%)	41	7 (0.10%)
2	496 (6.80%)	22	11 (0.15%)	42	5 (0.07%)
3	632 (8.66%)	23	12 (0.16%)	43	11 (0.15%)
4	201 (2.75%)	24	7 (0.10%)	44	5 (0.07%)
5	120 (1.64%)	25	11 (0.15%)	45	5 (0.07%)
6	161 (2.21%)	26	11 (0.15%)	46	1 (0.01%)
7	82 (1.12%)	27	10 (0.14%)	47	7 (0.10%)
8	67 (0.92%)	28	5 (0.07%)	48	1 (0.01%)
9	73 (1.00%)	29	6 (0.08%)	49	1 (0.01%)
10	37 (0.51%)	30	5 (0.07%)	52	2 (0.03%)
11	111 (1.52%)	31	9 (0.12%)	53	1 (0.01%)
12	53 (0.73%)	32	5 (0.07%)	54	1 (0.01%)
13	37 (0.51%)	33	3 (0.04%)	55	4 (0.05%)
14	40 (0.55%)	34	3 (0.04%)	56	3 (0.04%)
15	31 (0.42%)	35	4 (0.05%)	57	4 (0.05%)
16	14 (0.19%)	36	5 (0.07%)	58	1 (0.01%)
17	26 (0.36%)	37	4 (0.05%)	63	1 (0.01%)
18	13 (0.18%)	38	4 (0.05%)		
19	7 (0.09%)	39	6 (0.08%)		

Referring to the frequency table in Table 4.7.2, we observed that 52.84 % of the students did not yield any missingness. In contrast, 2,895 students yielded missing values in 1 to 10 variables. These students were considered as possessing a small number of missing values. The worst case was a student who yielded missing values in 63 out of 68 variables. Additionally, there were 44 students who yielded missing values in 31 to 40 variables, 43 students who yielded missing values in 41 to 50 variables, and 16 students who yielded missing values in 51 to 60 variables. Since these 104 cases were included in our analysis, more imputation work was required for these cases.

In the next section, we discuss the missing proportion plots and the missingness box plots.

4.7.1.2 Missingness Plots for Working Data Set

Two types of diagrams were adopted to investigate the missingness and its relationships within the working data set. These diagrams were outlined as:

1. The aggregate missingness pattern plot, which was for investigating the missingness of the drug-trying response variables, is displayed in Figure 4.3.

2. The missingness matrix plot, which depicted the missingness and levels of one variable against the missingness and levels of the other variables.

Figure 4.3: Aggregate Missingness Pattern Plot of Drug Responses. Figure 4.3(a) (Left) depicts the missing proportion of 15 drug responses; Pattern diagrams in Figure 4.3(b) (Right) depict the existence of missing value pattern among drug responses. Light blue represents observed cases, and red represents missing cases.

From Figure 4.3(a), the bar plots depicted the missingness of the drug responses in the working data set to be between 1.5 % and 2.5 %, resembling the missingness figures for drug responses from the original data set analysis as described in Chapter 3. From Figure 4.3(b), although most cases contained no missingness within any of the 15 drug-trying response variables, a large number of patterns that contained more than half of missingness were observed.

Following the discussion of missing proportion plots and the missingness pattern plots in Section 4.7.1.2, the investigation of the missingness of the working data set continued with missingness matrix plots. In the missingness matrix plots in Figures 4.4 to 4.5, all cases were sorted according to a particular variable, where values were highlighted by a grey scale that ranged from white to black, representing low to high levels of any variable. Missingness was marked in red. The purpose of the missingness matrix plots was to investigate, at all levels of a sorting variable, the pattern of missingness, thus diagnosing whether missingness of another variable depended on the sorting variable (e.g. whether missingness of cigarette smoking status depended on the attitude of the family on smoking).

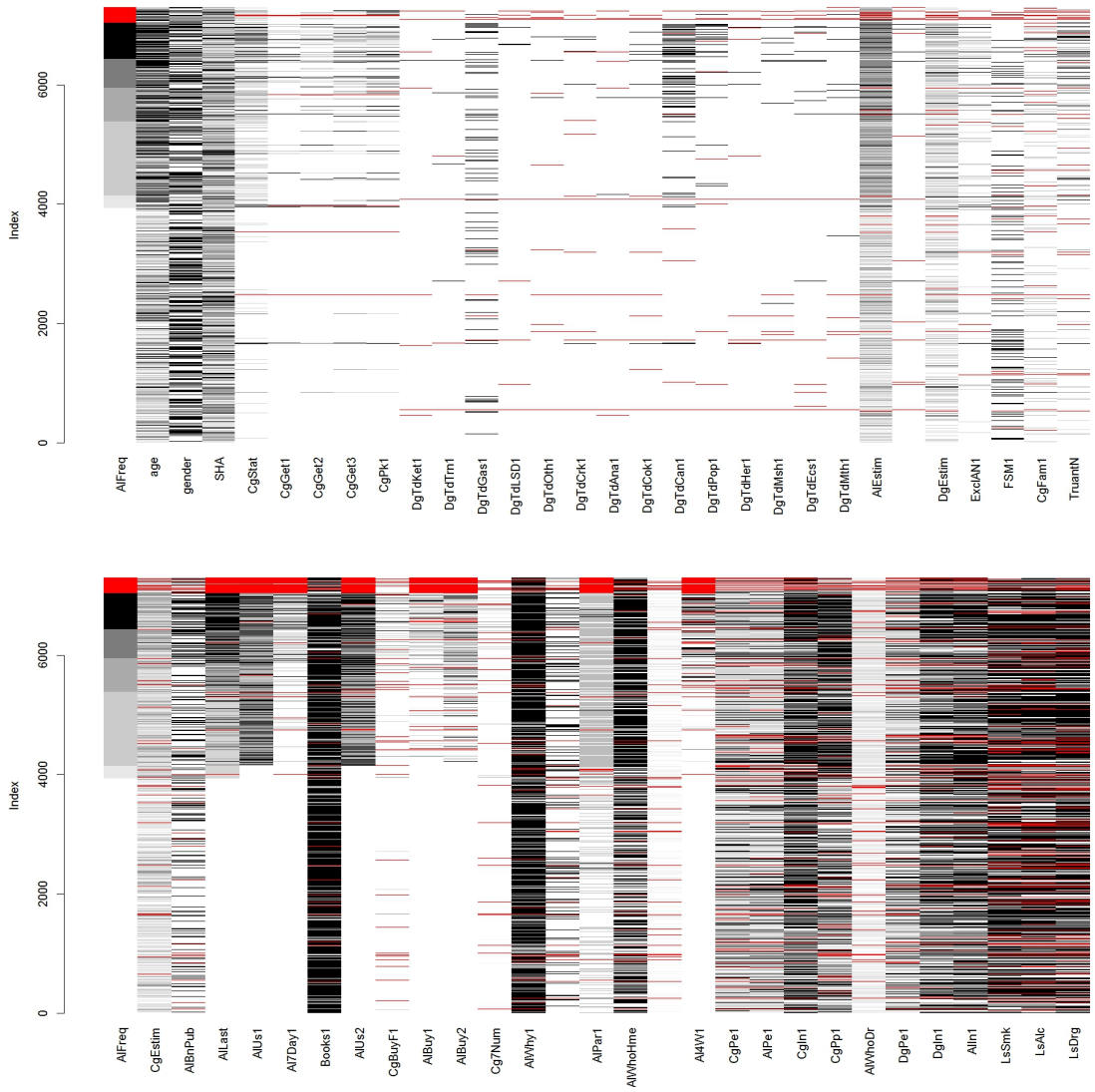

Figure 4.4: Missingness Matrix Plot of All 58 Covariates, Sorted by AlFreq Variable. Greyscale indicates levels from low(white) to high(black), red indicates missing values. More missingness on other covariates is observed for higher levels of the AlFreq variable, and the most missingness on other covariates is observed for the missing cases of the AlFreq variable.

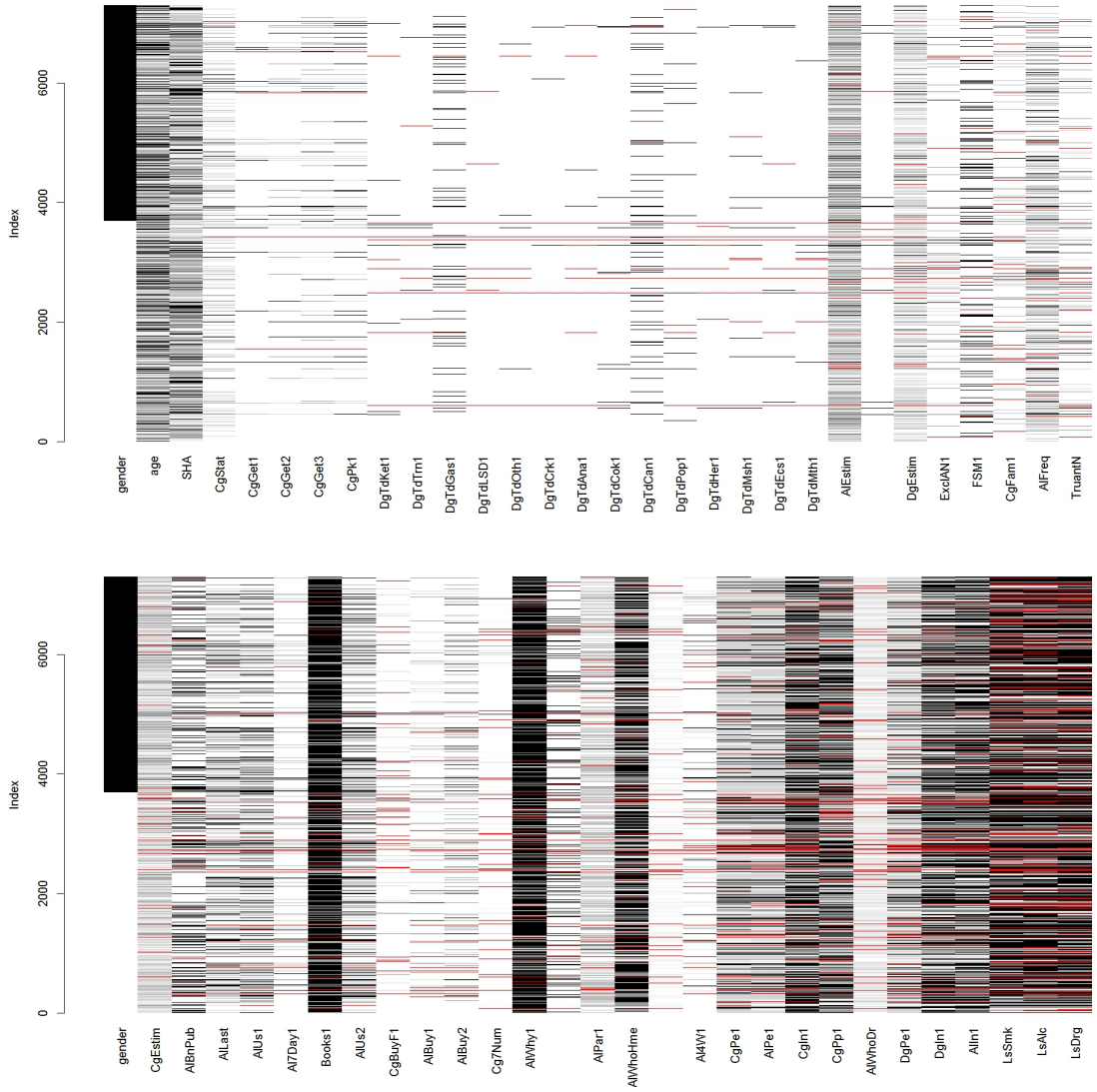

Figure 4.5: Missingness Matrix Plot of All 58 Covariates, Sorted by Gender Variable. Greyscale indicates levels of boys(white) and girls(black) for gender, and from low(white) to high(black) for other covariates, red indicates missing values. In this figure, boys yielded more missingness than girls.

The missingness matrix plots in Figure 4.4 indicated that for a particular student, when frequency of drinking (AlFreq) variable was missing, most variables appeared to be missing. This was due to the questionnaire design of the survey. From Figure 4.5, it was observed that more boys were associated with missing values in other variables than girls. From Figures 4.4 to 4.5, it was observed that missingness in the data set was related to several smoking, drinking and drug-related socio-demographic variables, for example, CgStat, AlFreq, Gender and Age, rendered the missingness in the data set to be missing at random (MAR).

To investigate the connection of each response variable in the working data set, which contained missingness in other covariates, we employed two criteria: (1) if missingness of a response variable depended on any other covariates and (2) if a response variable depended on any other covariates. In the following paragraphs, we defined a drug-trying response variable to be *dependent* on a covariate if either of the above two criteria, or both, held for such a drug-trying response variable.

The investigation of missingness of a drug-trying response variable on other covariates involved the Ridout's Test (Ridout and Diggle, 1991), which was explained in Section 4.3 above.

The significance threshold of 0.20 was adopted in this analysis for the purpose to include more potentially related variables in the regression model. The significance threshold of 0.20 has been suggested by Pearson (1938). The set of results corresponding to the 0.20 threshold are displayed in Figures 4.6 to 4.8. The missingness indicator plot is displayed in Figure 4.6, whereas the covariate significance plot is displayed in Figure 4.7, and the covariate dependency plot is displayed in Figure 4.8.

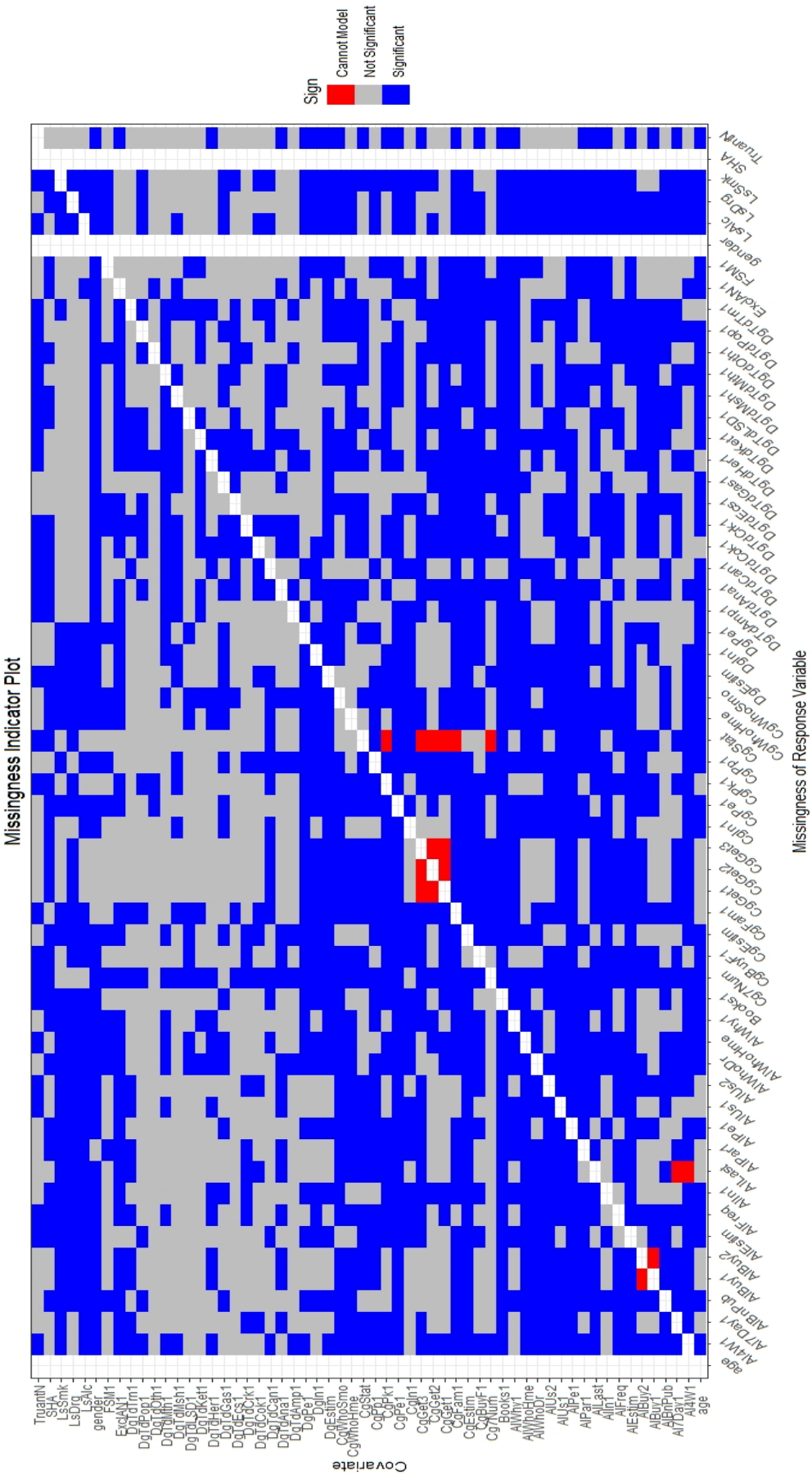

Figure 4.6: Missingness Indicator Plot of Significant Covariates at 0.20 Threshold. Blue represents significance at 20% level, grey represents insignificance, red represents errors in modelling

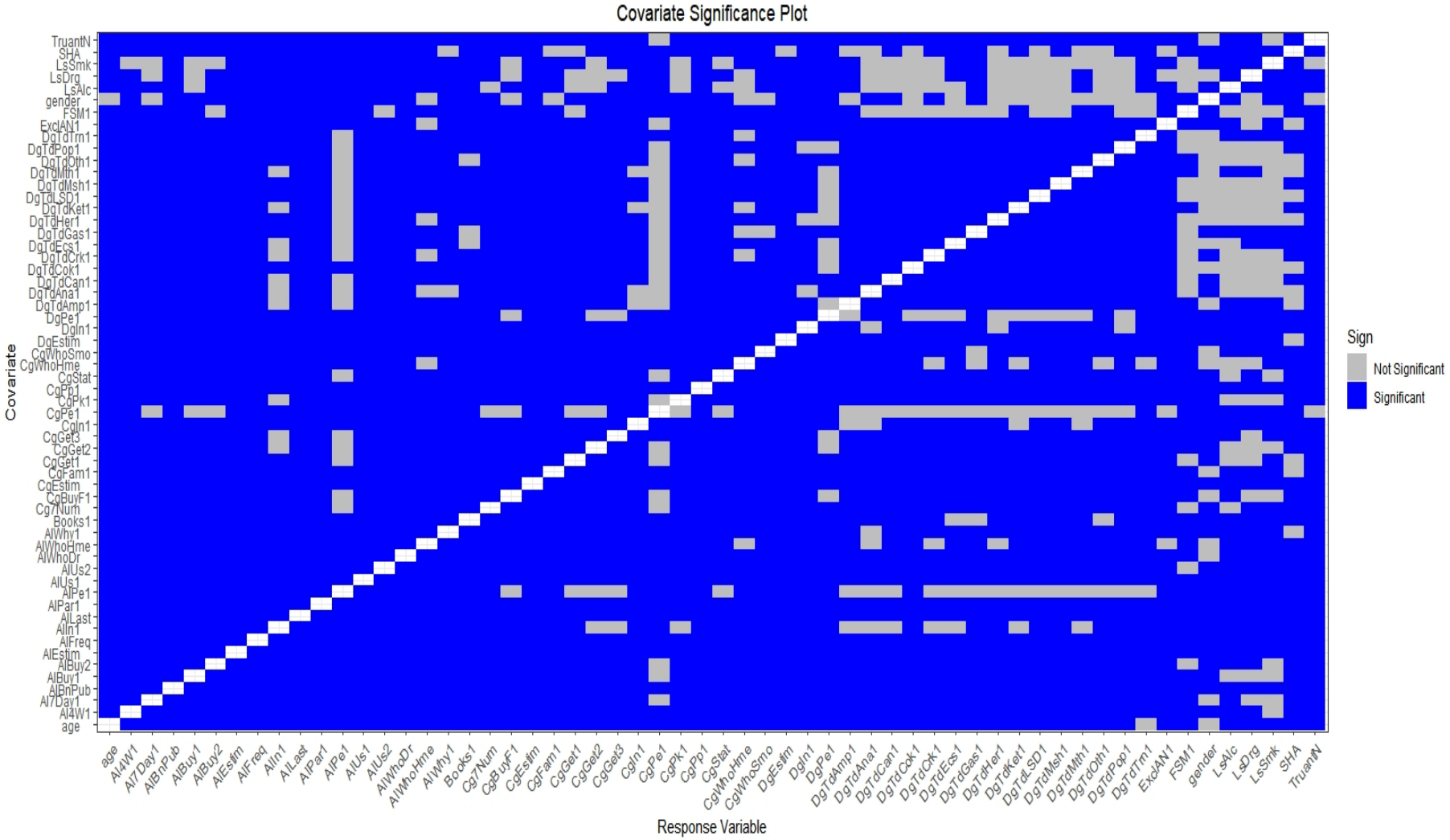

Figure 4.7: Covariate Significant Plot of Significant Covariates at 0.20 Threshold. Blue represents significance at 20% level, grey represents insignificance

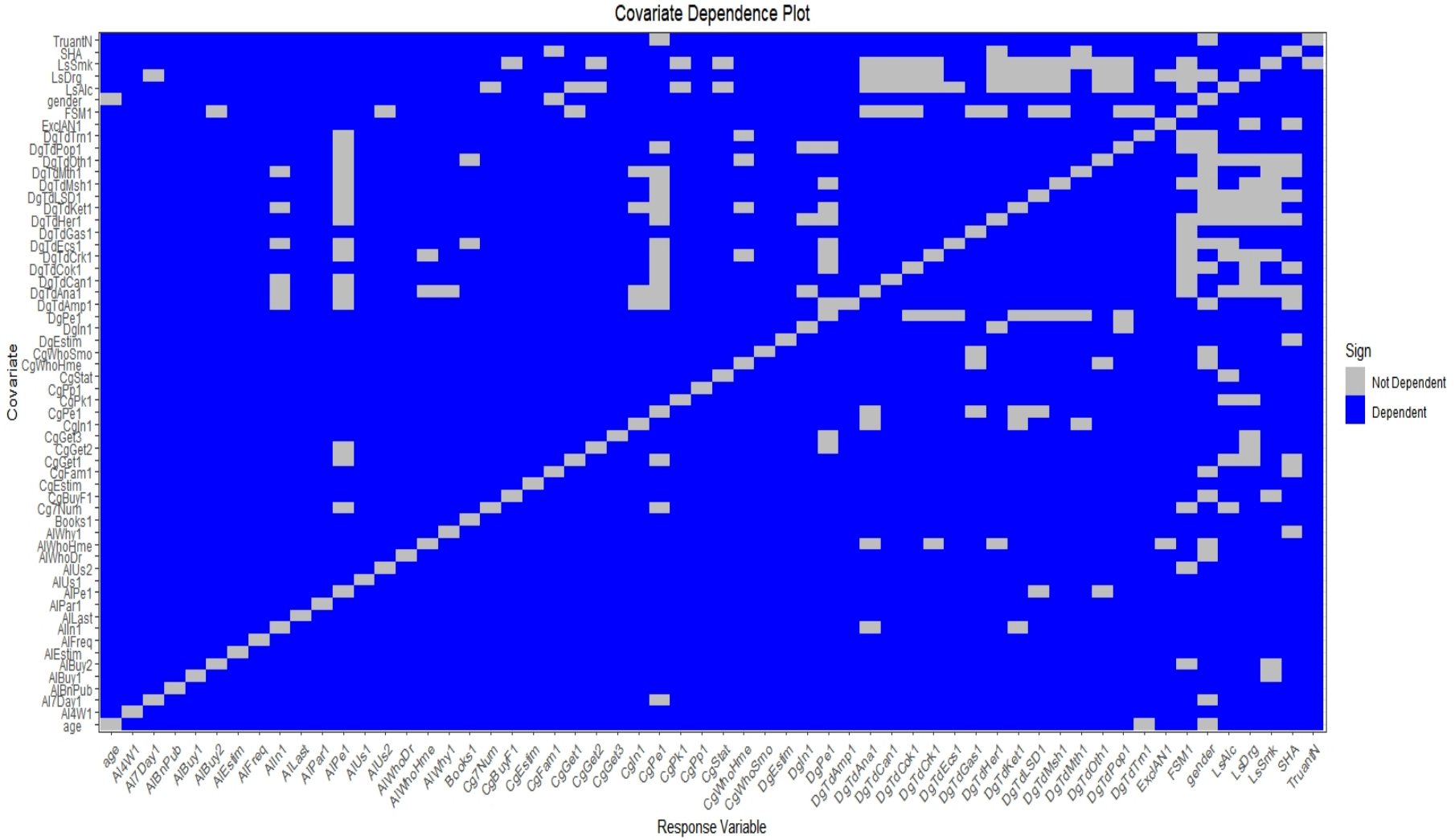

Figure 4.8: Covariate Dependence Plot of Significant Covariates at 0.20 Threshold. Blue represents significance at 20% level, grey represents insignificance

According to the missingness indicator plots in Figure 4.6, a few logistic regression models could not be fitted due to sparse cases of missingness in the complete set of individuals, thus leading to a slightly difficult investigation in missingness. However, from Figure 4.6, it was observed that the missingness of each of the 58 variables depended on at least 20 variables. The covariate significance plots in Figure 4.7 outlined the situation where all variables were associated with most of the other covariates (at least 30 out of 57 other covariates). The combined dependency indicator plots in Figure 4.8 generally depicted that all variables depended on most of the other covariates, supporting the belief that the missingness is MAR, and that all variables should be included in the MICE imputation, where all imputed variables are predicted based on all other covariates.

After identifying characteristics of the working data set throughout this exploratory data analysis of missing values, imputation of the missing data was then carried out on the working data set with missing values. The method of imputing the missing data is discussed in Section 4.7.2 below.

4.7.2 Building a Model for Imputation (only for FCS)

In this section, we discuss the plausibility of missing at random (MAR) assumption, followed by the procedures of imputing the missing data, the frequency tables and the polychoric correlation plots on imputed data sets. Firstly, the setup of the imputation of the missing data is discussed. Secondly, the multiple imputation by chained equations is considered for multiple imputation of the working data set, and their processes are described as well. Thirdly, proportion tables for drug-trying response variables and the polychoric correlation plots are interpreted.

4.7.2.1 Plausibility of Missing at Random Assumption and Ignorable Missingness Assumption

Referring to Section 1.3, the questionnaires of the "Smoking, Drinking and Drug Use Among Young People in England" survey series were conducted in an anonymous manner, i.e. none of the students' names or participating schools were recorded and reflected in the data set. Furthermore, the students were informed that their answers would be completely confidential. The students were also informed that only researchers could use the collected data for data analysis, and no such data would be revealed to any other personnel, such as police and members of the authorities, and therefore answering questions about smoking, drinking and drug use honestly would not be risky to them. In addition, during the Year 2010 survey, the researchers and staff had attempted strict survey procedures to keep the survey confidential and to raise the response percentage. Under the aforesaid confidential ways to collect data, the researchers and staff were expected to capture as many responses as possible, instead of attempting to relate any variable to missingness. Therefore, it could be reasonably assumed that the missing mechanism of the data set was ignorable and that it was MAR as suggested by Allison (2001). In other words, we could assume

that the missingness of any variable in the Year 2010 survey was not affected by the value of such variable itself, since we could assume that when making choices for questions in the questionnaire, the students were not affected by their worries that they would expose themselves into any risk. The assumption that the working data set was MAR was supported by the Missingness Matrix Plots shown in Figures 4.4 and 4.5 and Missingness Indicator Plots shown in Figures 4.6 to 4.8 as discussed in 4.7.1.2.

Moreover, the hypothesis test by Ridout and Diggle (1991) revealed that the MCAR hypothesis of the working data set was rejected at 5 % significance level, indicating that the working data set was not MCAR. Furthermore, Buuren (2012) stated that if MAR holds, for "practical purposes", the "missing data model" can be considered reasonable. Given the above-mentioned reasons, we could reasonably assume that the working data set was MAR and was ignorable.

After determining the working data set was MAR and was ignorable, we applied a suitable imputation called multiple imputation by chained equations (MICE) to the working data set, which is discussed in Section 4.7.2.2 below.

4.7.2.2 Variable Selection for Imputation

The first step of the MICE imputation setup was to select essential variables that covered all essential information of the working data set. In other words, derived variables and nested variables (variables that contained reduced levels from the original variables) were excluded from the imputation.

Referring to the working data set, the six variables, cigarette smoking status (CgStat1), smokers in house and where (CgWho1), types of sources of obtaining cigarettes usually (CgGet), frequency of drinking alcohol (AlFreq2), how respondents usually obtain alcohol (AlBuy), and drinkers in house and where

(AlWho1), were not included in the MICE. This was because these six variables were combined from their respective variables, as listed in Table 4.7.3 below:

Table 4.7.3: Table of Derived Variables in the Working Data Set

Combined variable	Original Variable
CgStat1	CgStat, CgIreg, Cg7Num
CgWho1	CgWhoSmo, CgWhoHme
CgGet	CgGet1, CgGet2, CgGet3
AlFreq2	AlFreq, Al7Day
AlWho1	AlWhoDr, AlWhoHme
AlBuy	AlBuy1, AlBuy2

We excluded these derived variables due to the following reasons: (1) since the levels in derived variables were well represented by particular combinations of original variables, levels in derived variables were redundant for imputation, and (2) high correlations might occur between derived variables and original variables.

Four nested variables were excluded because all nested variables exhibited very high correlations with parent variables, to an extent that singularities happened when both nested and parent variables were included in a regression model. Such nested variables are listed in Table 4.7.4 below.

Table 4.7.4: Table of Nested Variables in the Working Data Set

Nested variable	Parent Variable
CgNow	CgStat
AlEvr	AlFreq
Truant1	TruantN
ExclA1	ExclAN1

Although nested variables contained slightly more complete cases than parent variables, the difference in the number of complete cases was so small that it did not affect the superiority of parent variables over nested variables in providing valid information about a variable.

The second step of the setup was to specify the variable type (i.e. categorical (factor) or linear) for all variables in the working data set. Specifying a correct variable type was essential, especially for multiple imputation by chained equations, where generalized linear models were used for updating missing values.

In the imputation process, we made as few assumptions on each variable as possible. If a variable could be treated as either a nominal, ordinal or linear variable, we treated such variable as a nominal variable. This was because ordinal variables were subjected to an additional assumption that the odds of trying a certain drug increased when the variable level increased. Also, linear variables were also subjected to an extra assumption that the increase in the odds was constant between adjacent levels. However, we did not need to make these assumptions for nominal variables. Thus, treating a variable as nominal required the least assumptions to the variable.

The table describing the type of variables is presented in Appendix A.2. In general, there were four variables which were treated as numeric: (1) Cg7Num; (2) CgWhoSmo; (3) AlWhoDr and (4) Age. In this section, we only discuss the MICE, which was considered for imputation of the working data set.

In the MICE imputation, we adopted the `mice` package in `R` program. Seed number 4321 was adopted for all the MICE processes. The MICE imputation on the working data set with 58 variables was processed on Adelie Processor Cluster of Penguin Supercomputer Cluster at Lancaster University.

For a data set with all 58 variables, each variable depended on the other 57 variables during the MICE imputation. For a data set with 15 drug-trying response variables, each drug-trying response variable depended on the other 14 drug-trying response variables. Before imputation, all variables among each data set were sorted, from the first variable to be considered to the last, by ascending missing proportions.

4.7.3 Imputation Diagnostics/ Validation

During the 200 cycles of the MICE on both data sets, the variables, namely CgStat, CgGet3, CgPk1, AlFreq, AlLast, AlUs1, AlUs2, Al7Day1, AlPar1, AlWhoHme, LsSmk, LsAlc and LsDrg showed trends of changing estimates on trace plots of their mean and standard deviation at the initial stage of imputation. However, all variables were observed to converge after 150 imputations. The convergence plots for each of the variables in both data sets are presented in Figures 4.9 to 4.11.

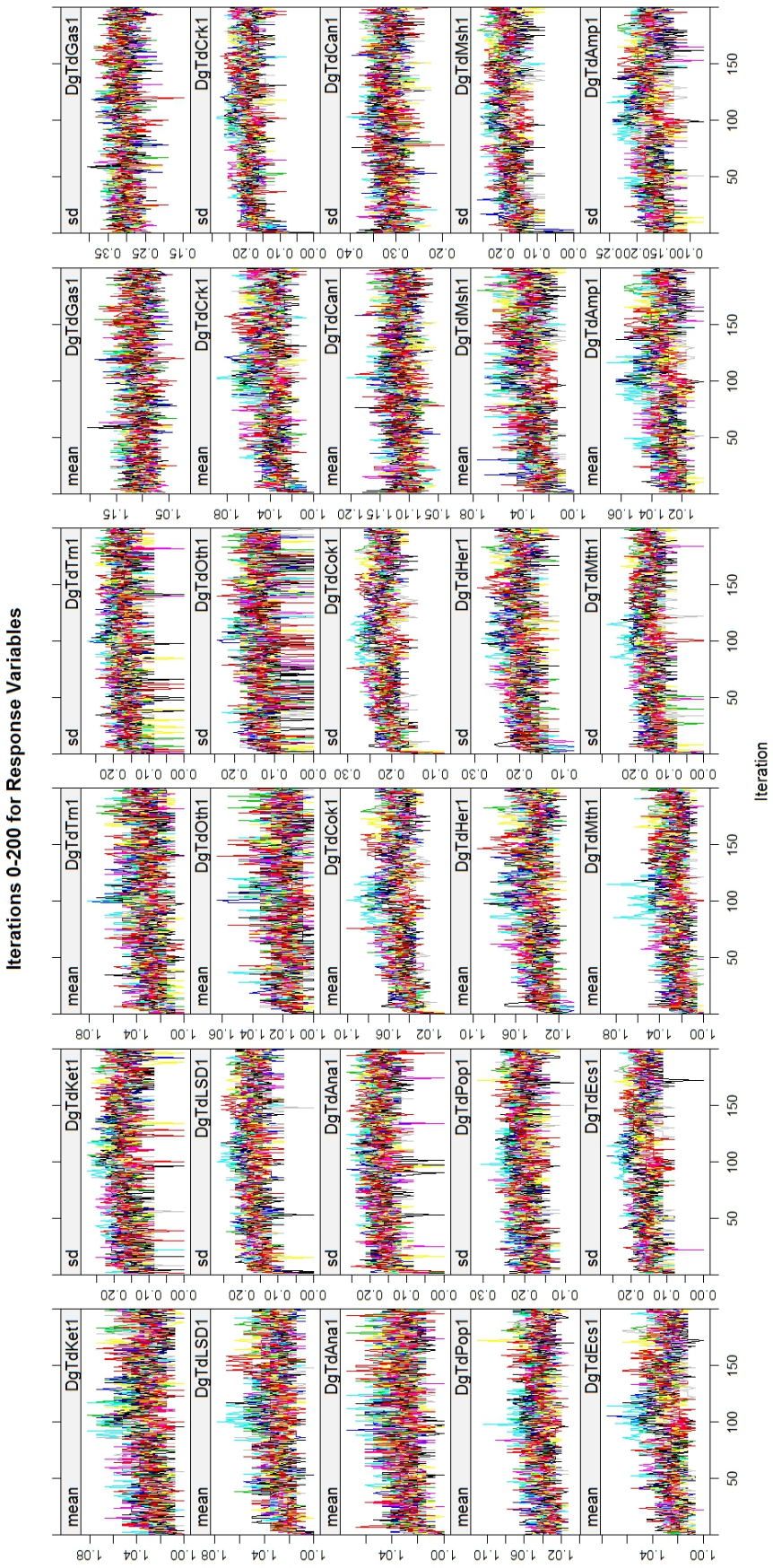

Figure 4.9: Convergence Plots for Ten Imputed Data Set under MICE Scheme 1 with Drug-trying Response Variables Only. All response variables converged after 200 iterations.

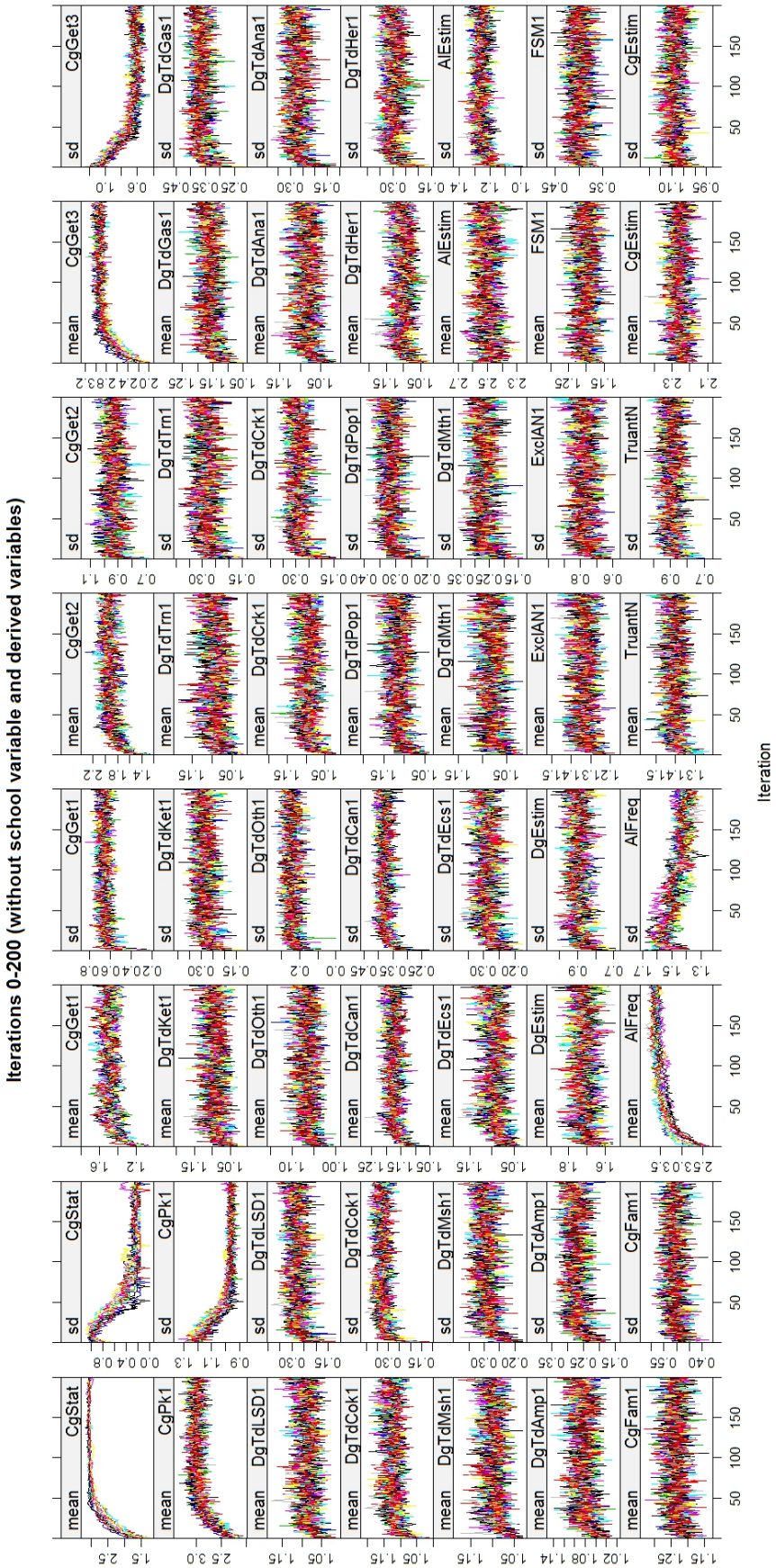

Figure 4.10: Convergence Plots for Ten Imputed Data Set under MICE Scheme 2 (Plot 1). All covariates converged after 200 iterations.

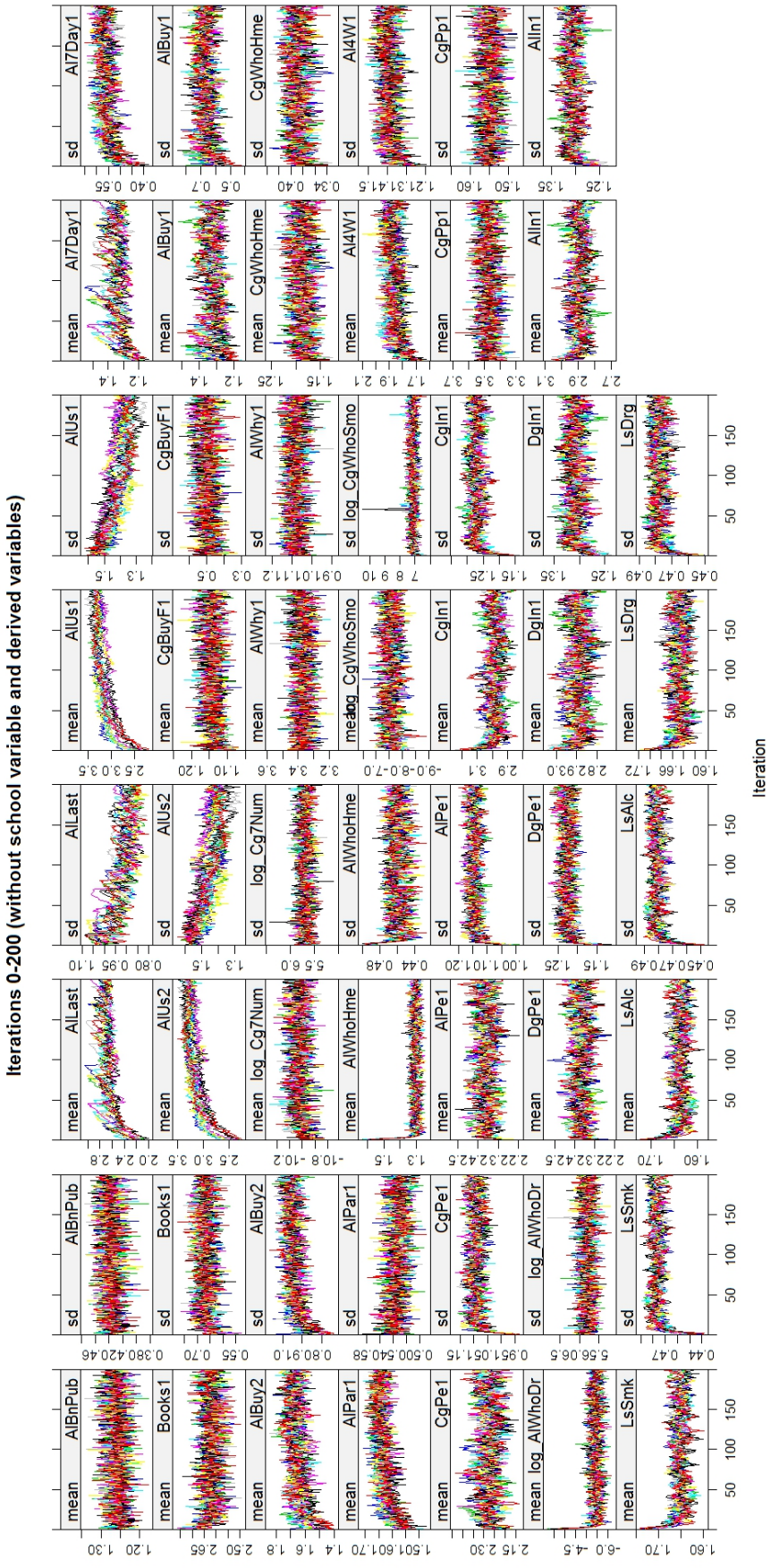

Figure 4.11: Convergence Plots for Ten Imputed Data Set under MICE Scheme 2 (Plot 2). All covariates converged after 200 iterations.

After the imputation of the working data set through the MICE scheme, we produced proportion percentage tables for drug-trying response variables of the imputed working data sets under two MICE imputation schemes, in order to compare with those of the original working data set and investigate the difference in the proportion of students trying a specific drug, as well as how the imputation from the two MICE schemes differed from each other. The related proportion percentage table is presented in Table 4.7.5.

Table 4.7.5: Proportion Percentage Table of Drug Response Variables in Original Working Data Set and Imputed Working Data Sets by the MICE Imputation (proportions in %) ("Org" means original working data set). Percentages are calculated without missing data.

		MICE Imputation Scheme 1									
Data Set	Org	1	2	3	4	5	6	7	8	9	10
Cannabis	9.24	9.27	9.17	9.27	9.29	9.29	9.33	9.28	9.32	9.29	9.27
Heroin	0.50	0.63	0.60	0.59	0.60	0.60	0.60	0.56	0.59	0.63	0.56
Cocaine	1.22	1.32	1.26	1.29	1.29	1.30	1.32	1.26	1.32	1.33	1.29
Magic Mushrooms	1.53	1.56	1.56	1.58	1.55	1.59	1.58	1.54	1.55	1.58	1.56
Crack	0.63	0.71	0.66	0.71	0.69	0.71	0.74	0.67	0.73	0.75	0.69
Methaone	0.73	0.77	0.74	0.77	0.75	0.81	0.78	0.73	0.77	0.78	0.77
Ecstasy	1.12	1.19	1.12	1.15	1.15	1.18	1.19	1.11	1.17	1.19	1.17
Amphetamines	0.94	0.97	0.96	1.00	0.97	1.03	1.00	0.96	1.00	1.01	0.97
LSD	0.59	0.63	0.59	0.64	0.62	0.66	0.63	0.62	0.62	0.66	0.63
Poppers	2.30	2.33	2.29	2.36	2.32	2.38	2.32	2.34	2.36	2.33	2.36
Ketamine	0.60	0.69	0.62	0.69	0.64	0.67	0.67	0.60	0.63	0.64	0.62
Anabolic Steroids	0.48	0.52	0.51	0.51	0.52	0.53	0.52	0.52	0.52	0.56	0.53
Gas	8.24	8.24	8.28	8.35	8.25	8.31	8.31	8.21	8.29	8.24	8.35
Other Drugs	0.46	0.48	0.47	0.51	0.47	0.45	0.49	0.49	0.48	0.49	0.49
Tranquillisers	0.45	0.51	0.45	0.53	0.48	0.47	0.48	0.51	0.48	0.49	0.51
		MICE Imputation Scheme 2									
Data Set		1	2	3	4	5	6	7	8	9	10
Cannabis		9.46	9.46	9.43	9.38	9.44	9.46	9.43	9.42	9.46	9.50
Heroin		0.85	0.74	0.71	0.64	0.73	0.67	0.70	0.77	0.75	0.79
Cocaine		1.52	1.41	1.44	1.38	1.51	1.47	1.38	1.45	1.45	1.48
Magic Mushrooms		1.70	1.67	1.70	1.70	1.75	1.70	1.70	1.69	1.74	1.73
Crack		0.82	0.85	0.85	0.81	0.82	0.75	0.77	0.75	0.81	0.84
Methaone		0.85	0.95	0.89	0.86	0.85	0.93	0.90	0.97	0.90	0.97
Ecstasy		1.33	1.23	1.27	1.21	1.33	1.30	1.29	1.38	1.32	1.34
Amphetamines		1.14	1.10	1.10	1.04	1.14	1.01	1.14	1.19	1.08	1.15
LSD		0.79	0.73	0.73	0.67	0.79	0.64	0.71	0.71	0.77	0.79
Poppers		2.54	2.49	2.43	2.56	2.56	2.40	2.44	2.54	2.56	2.51
Ketamine		0.73	0.75	0.78	0.73	0.74	0.78	0.71	0.81	0.78	0.78
Anabolic Steroids		0.71	0.71	0.55	0.60	0.58	0.56	0.69	0.70	0.67	0.70
Gas		8.50	8.36	8.46	8.48	8.42	8.31	8.44	8.43	8.43	8.48
Other Drugs		0.66	0.56	0.53	0.58	0.59	0.62	0.55	0.58	0.66	0.64
Tranquillisers		0.62	0.69	0.58	0.60	0.64	0.62	0.55	0.59	0.67	0.56

From Table 4.7.5, we observed similar proportion percentages for all drug-trying responses variables across all ten imputations for each imputation scheme. We found that most (over 90%) proportion percentages were inflated for all ten imputed data sets, from original data set with drug-trying response variables only (MICE Scheme 1). However, all proportion percentages were inflated for all ten imputed data sets with drug-trying response variables and smoking, drinking and drug-related socio-demographic covariates, which were imputed under MICE Scheme 2. This pattern was caused by drug-trying response variables being influenced by various smoking and drinking variables in their imputation models such as frequency of smoking (CgStat) and frequency of drinking (AlFreq). This highlighted how influential these smoking and drinking variables were on the proportion of the students trying every drug. In a similar way, those proportion percentages for data sets imputed under MICE Scheme 1 and MICE Scheme 2 were inflated from those for the original data set due to mutually positive association among drug-trying response variables.

4.8 Summary

Overall, we have identified missing data as a problem in our working data set. On average, there was approximately 4% of the data missing in each variable, with a range between 0.58% and 16.98%. Four variables did not contain any missingness. The highest value of missingness was found in questions relating to whether the students had taken any lessons about specific drug use and their effects.

When analysing any data with missingness, it was important to consider the underlying missing mechanism. In this chapter, we have introduced the missingness problem and defined missingness mechanisms and ignorability. Various exploratory methods were used to identify the missingness pattern and we de-

termined that the working data could be considered as missingness at random (MAR) and ignorable.

We considered and discussed two imputation methods, namely, the fully Bayesian framework and the multiple imputation by chained equations. The fully Bayesian framework had the advantage of being a one-stage method, when compared to the two-stage method of the MICE. However, the coding of the missingness model could be very complex under the fully Bayesian framework.

For the first stage of the MICE, we applied chained equations, which were similar with the Gibbs Sampler, on a fixed number of imputed data sets. Diagnostic plots showed convergence of these chains for all variables generally, though there were slow convergence for several variables. After multiple imputation by chained equations, the proportions of students trying certain drugs were similar across all ten imputed data sets, reflected by all 15 drug-trying response variables. For the second stage, analyses for a substantive model were performed on each of the imputed data sets. The estimates and covariance matrices among all imputed data sets were combined using Rubin's Rule. However, this required an analysis which produced a covariance matrix and a set of estimates for the substantive model.

For the rest of this thesis, we assume the working data set to be MAR, and we adopt the MICE for imputing all variables that contained missingness. Rubin's rule with Wald's test is adopted to test the significance of a covariance or an interaction term in regression models.

Chapter 5

Logistic Regression and Log-linear Analysis Models for Further Exploring Association and Interaction

5.1 Introduction

As discussed in Section 3.3.5, additional main findings from the exploratory data analysis of the working data set of this study provided hints and justification to further investigate the interactions among drug-trying response variables, the smoking, drinking and drug-related socio-demographic variables. In this chapter, generalized linear models are applied to further explore possible interactions among the binary drug-trying response variables in the working data set of this study and to understand more the associations of the smoking, drinking and drug-related socio-demographic covariates with drug-trying response variables. The first type of model applied is the univariate logistic regression model, a type of generalized linear model (GLM). In the univariate logistic regression analysis model, a single binary drug-trying response variable is modelled against

covariates and other drug-trying response variables. The univariate logistic regression model is repeated for each of the 15 drugs. The second type of model applied is the log-linear analysis model, which is another type of GLM, in which the frequencies of students in all combinations of the 15 drug-trying response variables are modelled against the main effects and the first order interactions among the drug-trying response variables.

In this chapter, firstly, a brief introduction to the univariate logistic regression model is made and then each drug-trying response variable is modelled against all covariates and other drug-trying response variables. In the analysis, a backward elimination procedure is adopted to eliminate covariates with little explanatory value. In the backward elimination procedure, each drug-trying response variable is regressed against the other drug-trying response variables and all other explanatory covariates, i.e. those smoking, drinking and drug-related socio-demographic covariates. Therefore, in the univariate logistic regression analysis, one-way interactions between one drug-trying response variable and the smoking, drinking and drug-related socio-demographic covariates as well as other drug-trying response variables is examined.

Secondly, a brief introduction to the log-linear analysis model is made. This is a type of Poisson GLM where the counts of each combination of the 15 drug-trying response variables are modelled against the main effects and the first order interactions among the drug-trying response variables to identify significant two-way interactions of these variables. We again employ Rubin's rule and apply backward elimination in running the model.

Finally, we compare the interactions found using the univariate logistic regression model with those found using log-linear analysis model and discuss the results.

5.2 Univariate Logistic Regression Model

5.2.1 Introduction

The main aims of conducting the univariate logistic regression model are two-fold:

1. To investigate the relationship of every drug-trying response variable with the smoking, drinking and drug-related socio-demographic factors, along with other drug-trying response variables.

2. To serve as a useful guide for variable selection in a latent class analysis (which will be discussed in Chapter 7).

5.2.2 Theory

There are two common characteristics of univariate generalized linear models (Dobson and Barnett, 2008):

1. The distribution describing the dependent variable is from the exponential family.

2. Let the mean of each response i be $\mathbf{E}(y_i) = \mu_i$ and denote the monotone link function to be either $g(\mu_i)$ or η_i, which relates μ_i to the linear predictors x_i with a set of parameters β.

$$\eta_i = g(\mu_i) = x_i^T \beta.$$

In modelling a binary variable (e.g. a drug-trying response variable) using the univariate logistic regression model, an appropriate link function is a logit(μ_i) link. For respondents $i, i = 1, \ldots, N$, the probability of a positive response for

respondent i given the predictors x_i is denoted by $p_i = P(y_i = 1 \mid x_i)$, where y_i denotes the response.

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(p_i) \\
\mathbf{E}[y_i] &= p_i \\
\text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta .
\end{aligned}
$$

The log-likelihood for the univariate logistic regression model is expressed by the following equation:

$$
\ell(\mathbf{p};\mathbf{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right].
$$

(McCullagh and Nelder, 1999)

The likelihood for the univariate logistic regression model is maximized by the repeated use of the Iteratively Weighted Least Squares (IWLS) algorithms. The IWLS function estimates the mode value of the log-likelihood and uses a local quadratic approximation to the log-likelihood function to estimate the variance. The IWLS is the repeated application of the Newton-Raphson method (Hazewinkel, 1994). Details of the IWLS algorithms can be found at Green (1984).

In this research, each set of parameters in each univariate logistic regression model with each imputed data set respectively were estimated by the maximum likelihood function, which was maximised through the IWLS algorithm. These estimated parameters for all corresponding imputed data sets were combined by using Rubin's rule. Backward elimination was employed on the M imputed data sets, which involved the combination of Rubin's rule (Rubin, 1987) and Wald's test (Wald, 1945), to eliminate the covariates one by one in each backward elimination step in order to reach the most parsimonious model. Rubin's

rule and Wald's test can be referred to Section 4.5.3.3 respectively. The procedure of backward elimination based on Rubin's rule is described below.

The backward elimination process begins by fitting saturated regression models to all M imputed data sets; the consequent estimates and standard errors of these M saturated models are then pooled and combined using Rubin's rule. Wald's test is then conducted for each estimate in the saturated model. If the combined p-value of an estimate is greater than 0.05, then the term is considered to be discarded from the model, otherwise, the variable is retained in the model.

At this iteration, only the term with the highest combined p-value is discarded from the saturated model. Afterwards, the M imputed data sets are fitted with a reduced regression model without the discarded term. The subsequent estimates and standard errors of these M saturated models are pooled and combined by Rubin's rule. Wald's test is consequently conducted and the insignificant term at 5% significant level with the highest combined p-value at this stage is discarded from the reduced model. This process repeats with one term discarded at each iteration until no insignificant terms remained in the model. Such model at this step is considered as the final model.

For complete case analysis, Wald's test is adopted as the term selection test for the backward elimination.

5.2.3 Application of Univariate Logistic Regression Model

In this research, the univariate logistic regression model predicted the students' drug-trying behaviour for each drug-trying response variable with respect to the smoking, drinking and drug-related socio-demographic covariates. Two groups of the univariate logistic regression model were employed:

Model 1: univariate logistic regression models which consisted of other 14 drug-trying response variables as covariates;
Model 2: univariate logistic regression models which consisted of other 14 drug-trying response variables as well as the smoking, drinking and drug-related socio-demographic variables as covariates.

The model 1 was set up to investigate solely how the use of a drug was related to the use of other drugs, whereas the model 2 was set up to investigate how other drug-trying response variables, together with smoking, drinking and drug-related socio-demographic covariates, predicted the probability for trying each drug.

For model 1 (i.e. the univariate logistic regression models with 15 drug-trying response variables), two imputation schemes, namely scheme 1: MICE Imputation, FCS based on 15 drug-trying response variables only and scheme 2: MICE Imputation, FCS based on full data frame, were adopted for imputation of the data. Each imputation scheme generated ten corresponding imputed data sets.

When dealing with the ten imputed data sets, which were generated from each imputation scheme, two modelling processes were used: (1) the saturated model included all other 14 drug-trying response variables as covariates without backward elimination and (2) the final model resulting from backward elimination which began with all other 14 drug-trying response variables as covariates.

Backward elimination started with fitting each of the ten univariate logistic regression models with each of the ten corresponding imputed data sets in the `R` program through the `glm` function, which used the IWLS function for maximizing the likelihood of each model. The ten resulting sets of estimates and

standard errors, each from the pattern table of each imputed data set, were then pooled by Rubin's rule. Wald's test was then conducted for each covariate to determine whether it was significant at 5 % significance level. Among all insignificant terms found, the term with the highest p-value (denoted as X_{P1}) was discarded in each step.

Then ten imputed data sets were fitted again with the univariate logistic regression model without the X_{P1} term, and the whole process was repeated and the term with the highest p-value at this step, X_{P2}, was discarded from the model. The ten imputed data sets were then fitted again with the univariate logistic regression model without both X_{P1} and X_{P2} terms, and the same procedure was repeated for every step until no insignificant terms remained in the model. This ultimate univariate logistic regression model was the final model.

For comparison purposes, the above-mentioned modeling process that was applied to the imputed data sets were applied to complete case analysis situation for model 1 as well. In complete case analysis situation, Wald's test was adopted to test each covariate.

The results of the univariate logistic regression models involving 15 drug-trying response variables only (i.e. model 1) are discussed in Section 5.2.4 and the results of the univariate logistic regression models involving the smoking, drinking, drug-related socio-demographic covariates and other drug-trying response variables (i.e. model 2) are examined in Section 5.2.5. Each section commences with the discussion of the significant variable indicators, follows by the discussion of tables of estimates and standard errors, which are presented in Tables B.3.1 to B.3.5 for model 1 in Appendix B.3 and Tables B.4.1 to B.4.21 for model 2 in Appendix B.4 respectively. Finally, results of the univariate logistic regression results in respect of model 1 and model 2 are compared to investigate the effect

of including smoking, drinking and drug-related socio-demographic covariates in the univariate logistic regression analysis.

5.2.4 Univariate Logistic Regression Model with Other Drug-trying Response Variables as Covariates

In this section, we concentrate on investigating each binary drug-trying response variable, as a function of other drug-trying response variables via the univariate logistic regression models among fifteen drug-trying response variables only (i.e. model 1). As mentioned in Section 5.2.3, this analysis involved two modelling processes, namely the saturated model which included all 14 other drugs as covariates, and the final model resulting from backward elimination. In analysing the results of the univariate logistic regression model 1 here, both final models and saturated models were considered. The purpose of implementing the final models with backward elimination was to find the most parsimonious model for predicting students' drug-trying behaviour based upon drug-trying response variables. The purpose of implementing the saturated models without backward elimination was to provide compatible models for comparison.

5.2.4.1 Results of the Univariate Logistic Regression Model with Other Drug-Trying Response Variables as Covariates

For Model 1, we constructed a covariate sign plot for the final models with backward elimination conducted, which indicated the form of the relationship in Figure 5.3. In this research, the covariate sign plot is a grid plot displaying three colours for each combination of response variable and covariate variables under three groups: (1) positive associations, which are displayed in blue; (2) negative associations, which are displayed in red and (3) not significantly associated (with p-value larger than 0.05), which are displayed in grey. The covariate sign plot for the saturated models, without backward elimination conducted,

is shown in Figure 5.4, across data sets imputed under all two MICE schemes and complete case analysis. The related tables contained estimates and standard errors for final models and saturated models are shown in Tables B.3.1 to B.3.5 and Tables B.3.6 to B.3.10 respectively in Appendix B.3.

Before considering the results shown in Figure 5.3 and 5.4, it is worth to note that in Chapter 3 Section 3.3.4, Figures 3.3 to 3.5 in respect of polychoric correlation plots have already shown that generally all the 15 drug-trying response variables were strongly and positively correlated with each other. The log-odds ratio heat plots in Figures 5.1 and 5.2 and the covariate sign plots in Figures 5.3 and 5.4 further explore such relationships among the 15 drugs in a greater detail.

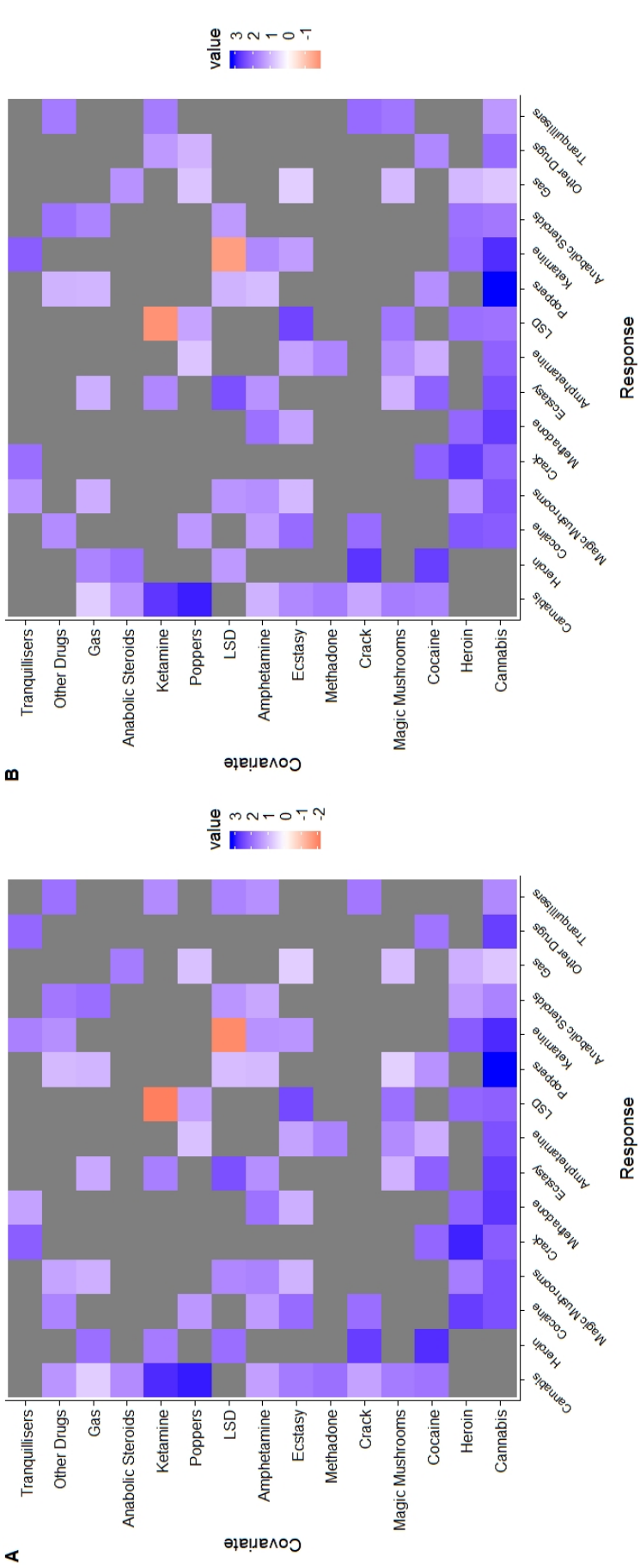

Figure 5.1: Log-odds Ratio Heat Plot of Logistic Regression Final Models under Scheme 1 (left) and Scheme 2 (right). Blue represents positive log-odds ratios, red represents negative log-odds ratios and grey represents insignificant or non-applicable log-odds ratios.

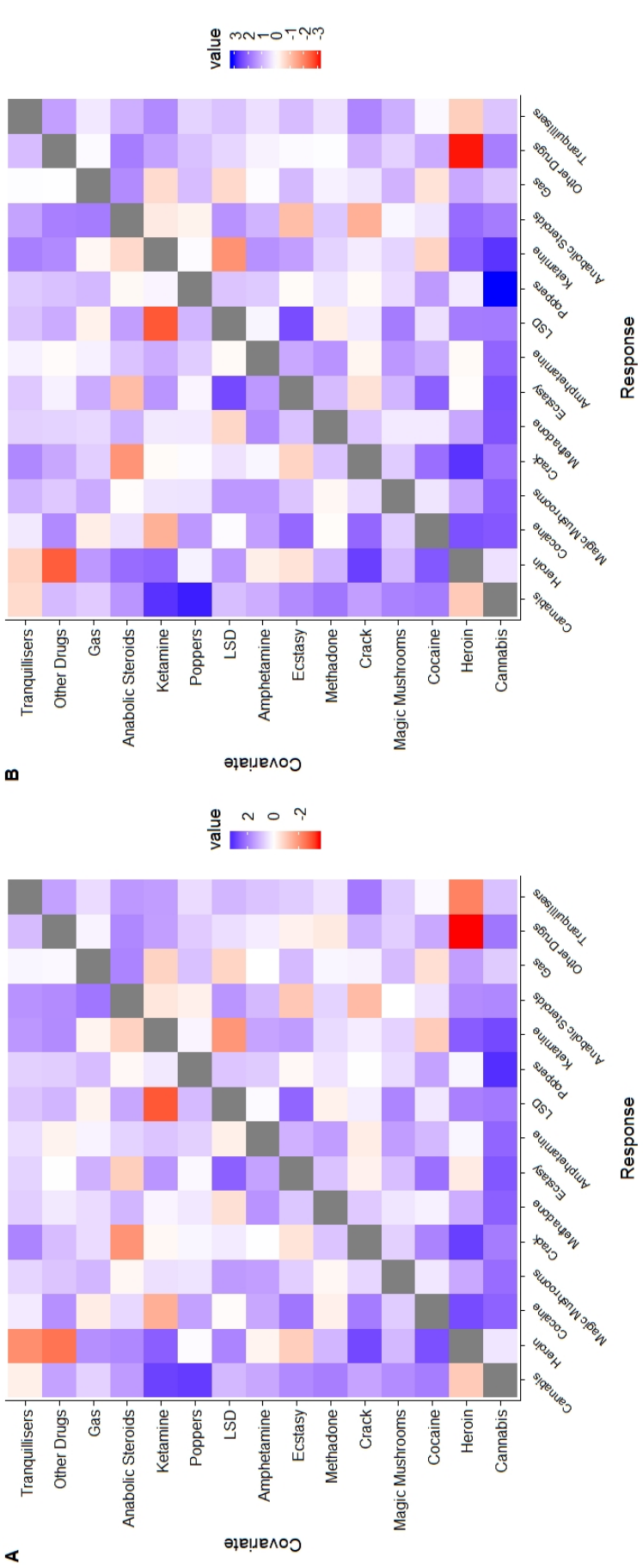

Figure 5.2: Log-odds Ratio Heat Plot of Logistic Regression Saturated Models under Scheme 1 (left) and Scheme 2 (right). Blue represents positive log-odds ratios, red represents negative log-odds ratios and grey represents insignificant or non-applicable log-odds ratios.

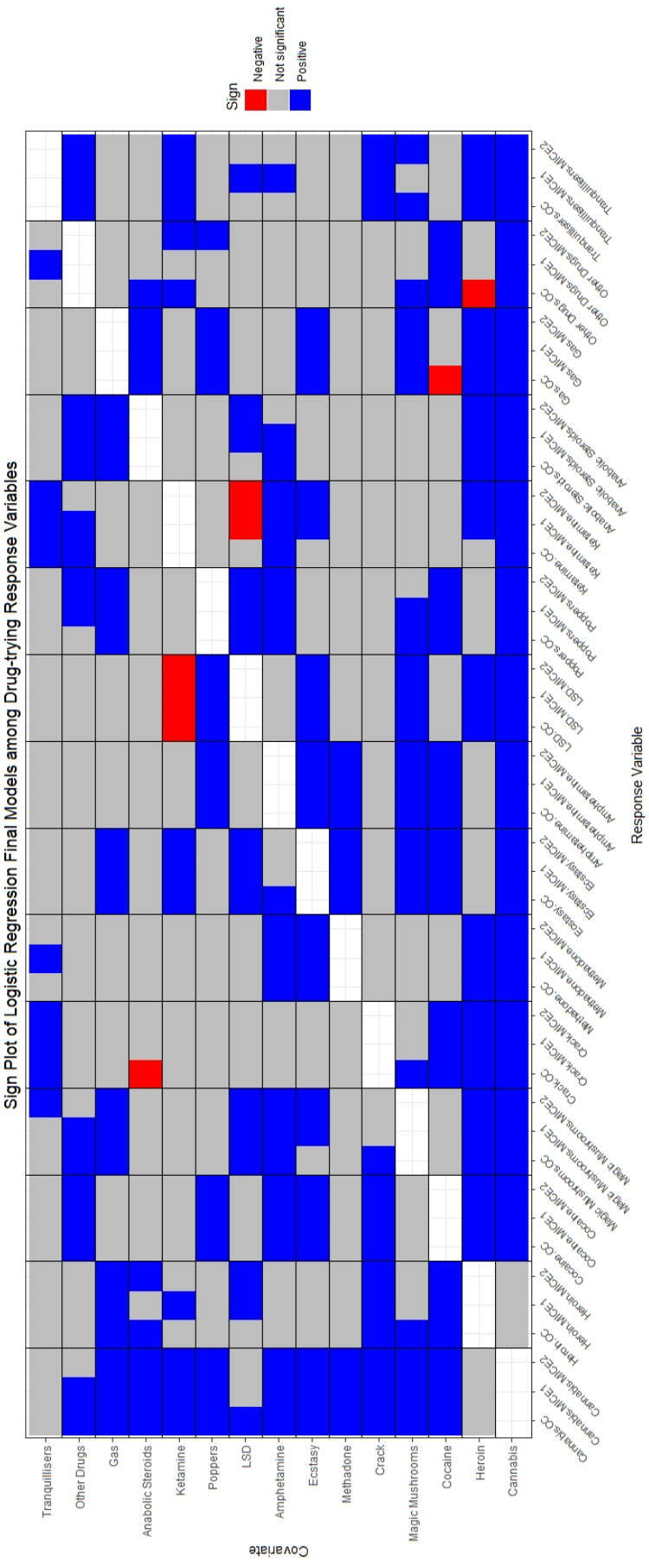

Figure 5.3: Covariate Sign Plot of Logistic Regression Final Models. (Axis label description - CC: Complete case analysis; MICE1: MICE Imputation Scheme 1; MICE 2: MICE Imputation Scheme 2) (Grid Colour Description - Blue: Positive Association; Red: Negative Association; Grey: No Significant Association at 5% significance level; White: Not applicable)

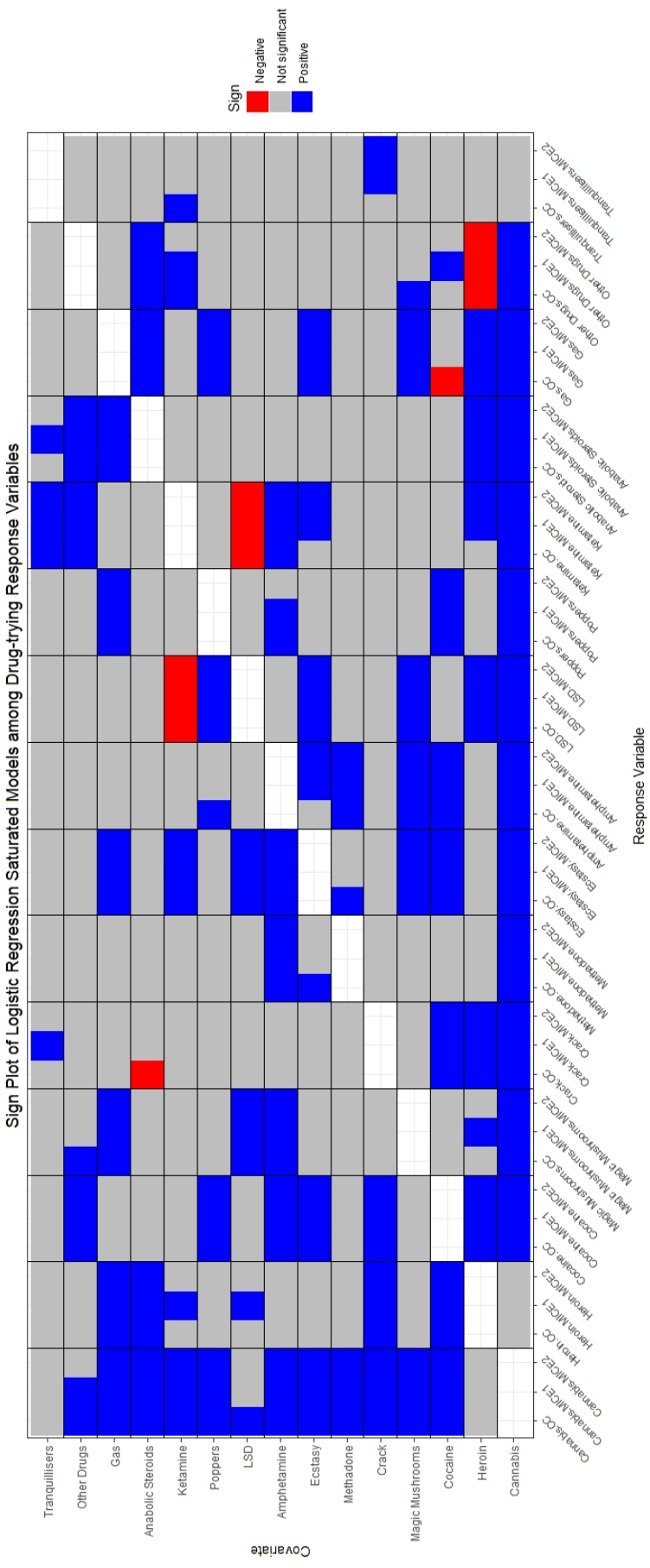

Figure 5.4: Covariate Sign Plot of Logistic Regression Saturated Models, (Axis label description - CC: Complete case analysis; MICE 1: MICE Imputation Scheme 1; MICE 2: MICE Imputation Scheme 2) (Grid Colour Description - Blue: Positive Association; Red: Negative Association; Grey: No Significant Association at 5% significance level; White: Not applicable)

Both log-odds ratio heat plots in Figures 5.1 and 5.2 showed that most remaining terms in the 15 final models were positive (log odds ratios larger than zero), and two imputation schemes generated similar results. The covariate sign plot of the univariate logistic regression final models with backward elimination, which is shown in Figure 5.3, showed that most remaining terms in the 15 final models possess positive associations, indicating that if a student has tried a specific drug, the student was more likely to try other drugs. The only exception was the relationship between LSD and ketamine. In addition, cannabis was found to associate with almost all drug-trying response variables except heroin. On the other hand, crack, methadone, other drugs and tranquillisers were associated with relatively a smaller number of other drug-trying response variables.

The covariate sign plot of the univariate logistic regression saturated models without backward elimination, which is shown in Figure 5.4, exhibited similar significance patterns as shown in Figure 5.3. The plot also shows fewer significant terms with a positive association between any pair of drug-trying response variables and one additional negative association between heroin and other drugs. The slight discrepancies between significant terms in the two sign plots might due to the adjustments of the covariate terms made through backward elimination process in the final models.

Referring to Figure 5.3, in most of the covariate terms, the results generated from imputed data sets by MICE schemes 1 and 2 generally agreed with each other. The slight discrepancies in the results between these two schemes might be due to slight differences in the percentages of students trying each of the 15 drugs, caused by the influence of the smoking, drinking and drug-related socio-demographic covariates. Table 4.7.5 in Section 4.7.3 illustrates such slight differences. These differences reflected the influence of the smoking, drinking and drug-related socio-demographic covariates on the results of the imputation

in the drug-trying response variables, and eventually the results of the univariate logistic regression final models among drug-trying response variables.

In terms of the numerical results of the univariate logistic regression final models in respect of 15 drug-trying response variables, Tables B.3.1 to B.3.5 in Appendix B.3 showed a general picture that almost all the 15 drugs were positively associated with each other. According to MICE scheme 2, cannabis was found to have positive associations with 10 drugs. Cocaine, magic mushrooms and ecstasy were shown to have positive associations with 7 drugs. Poppers, amphetamines and gas were displayed to have positive associations with 6 drugs. Heroin, tranquillisers and anabolic steroids were found to have positive associations with 5 drugs, as well as LSD and ketamine, but the latter two included a negative interaction relationship with one drug. Finally, methadone, crack and other drugs were shown to have positive associations with 4 drugs. As cannabis and gas were two drugs reported by a higher proportion of students who had tried them, as reported in the Year 2010 Survey, these two drugs would further be discussed.

We concentrated on discussing the two groups of univariate logistic regression model, as described in this section. Each univariate logistic regression model was analysed with a specific drug, namely cannabis, gas, crack or tranquillisers, as the response respectively, which was modelled with imputed data set under MICE scheme 2. Firstly, we discussed the two models with the highest proportion of students trying a specific drug: the logistic model with cannabis as the response variable and another with gas.

Focusing on the imputed data set under MICE scheme 2, Table B.3.1 in Appendix B.3 showed the probability for a student who had tried cannabis but without trying other types of drugs was at an odds ratio of $e^{-2.7497} = 0.06395$. The students who had tried cocaine were more likely to try cannabis, at an odds

ratio of $e^{1.7329} = 5.6570$. The students who had tried magic mushrooms were more likely to try cannabis, at an odds ratio of $e^{1.7882} = 5.9787$. The students who had tried crack, methadone, ecstasy, anabolic steroids or amphetamines were similarly more likely to try cannabis. In the other cases, the students who had tried poppers or ketamine were more likely to try cannabis, at odds ratios of $e^{2.9916} = 19.9175$ and $e^{2.7259} = 15.27019$ respectively. To a lesser extent, the students who had tried gas were more likely to try cannabis, at an odds ratio of $e^{0.6938} = 2.0013$. These estimates appeared to be similar with those corresponding estimates generated under saturated model, as seen from Table B.3.6 in Appendix B.3.

Similarly, from Table B.3.5 in Appendix B.3, the probability the students who had tried gas but without trying other types of drugs was $e^{-2.6213} = 0.07271$. The students who had tried cannabis were more likely to try gas at an odds ratio of $e^{0.7705} = 2.1608$, and the students who had tried heroin, magic mushrooms, ecstasy or poppers were more likely to try gas, at odds ratios of $e^{0.9839} = 2.6748$, $e^{0.9361} = 2.5500$, $e^{0.6683} = 1.9509$ and $e^{0.8132} = 2.2551$ respectively. In addition, the students who had tried anabolic steroids were more likely to try gas, at an odds ratio of $e^{1.49} = 4.4371$. All covariate terms, except the estimate of ecstasy, appeared to be similar with those estimates generated without backward elimination, as seen from Table B.3.10 in Appendix B.3. The discrepancy of the estimates of the ecstasy covariate term could be explained by the adverse confounding relationship of the students who had tried ecstasy and who also tried heroin, crack or anabolic steroids, of which the corresponding estimates were negative as shown in Table B.3.8 in Appendix B.3.

When comparing the results of the univariate logistic regression final model with cannabis to the results of the univariate logistic regression final model with gas, it was observed that the model with the cannabis as a response vari-

able yielded more significant terms and larger estimates than the final model with gas. This observation indicated a larger proportion of students who had tried cannabis and who also tried other types of drugs than those who had tried gas and other types of drugs. Therefore, it highlighted a stronger association of trying cannabis with trying other types of drugs.

After discussing cannabis and gas, of which a higher proportion of the students who reported trying them in the Year 2010 Survey, we continued to discuss the drugs with a moderate and the lowest proportion of the students who reported them in the Year 2010 Survey, namely crack and tranquillisers, in Tables B.3.1 and B.3.5 in Appendix B.3.

Focusing on the imputed data set under MICE scheme 2, Table B.3.2 in Appendix B.3 showed the probability of the student trying crack without trying other types of drugs was $e^{-6.255} = 0.001921$. Students who had tried cannabis were more likely to try crack, with an odds ratio of $e^{2.1344} = 8.4520$. Similarly, students who had tried heroin, cocaine or tranquillisers were more likely to try crack, with odds ratios of $e^{2.688} = 14.7022$, $e^{2.1769} = 8.8189$ or $e^{1.9876} = 7.2980$ respectively. These estimates, which were generated by the final models, appeared to be similar with those corresponding estimates generated by the saturated models, which could be referred to Table B.3.7 in Appendix B.3.

Finally, from Table B.3.5 in Appendix B.3, the probability of the students trying tranquillisers but without trying other types of drugs was $e^{-6.2404} = 0.001949$. The students who had tried cannabis were more likely to try tranquillisers, at an odds ratio of $e^{1.4099} = 4.0955$. Similarly, the students who had tried magic mushrooms, crack, ketamine or other drugs were more likely to try tranquillisers, at odds ratios of $e^{1.8991} = 6.6799$, $e^{2.0391} = 7.6837$, $e^{1.7976} = 6.0351$ or $e^{1.8307} = 6.2383$ respectively. All covariate terms generated in the final models, except for the

estimate of other drugs, appeared to be similar with those corresponding estimates generated in the saturated models, which are shown in Table B.3.10 in Appendix B.3. The discrepancy of the estimate of the other drugs covariate term could be explained by the adverse relationship of students who had tried other drugs and who had also tried heroin, of which the corresponding estimates were negative, as shown in Table B.3.10 in Appendix B.3.

In addition, according to Tables B.3.1 to B.3.5 and Tables B.3.6 to B.3.10 in Appendix B.3, the estimates of significant terms in the final models and those in the saturated models for the two sets of imputed data under MICE scheme 1 and scheme 2 were found to be quite similar. This finding was valid to any covariates that existed in the final models and the saturated models with several exceptions, such as the estimates of significant terms in the final and saturated models of: (1) LSD or other drugs covariates with heroin as the response variable; (2) tranquillisers covariate with ketamine as the response variable and (3) heroin or gas covariates with anabolic steroids as the response variable. Such differences were due to different imputed responses between two sets of ten imputed data sets under two different MICE schemes (i.e. MICE scheme 1 and MICE scheme 2) respectively.

The results of the univariate logistic regression models among 15 drug-trying response variables fitted on data sets, imputed through two MICE schemes, are compared with those fitted on data sets under complete case analysis, as shown in Tables B.3.1 to B.3.5 in Appendix B.3. The results of estimates and standard errors from the two MICE schemes applied on 15 drugs appeared to be closer in terms of their magnitudes. Those from the complete case analysis appeared to be farther in terms of their magnitudes from either set of results for the two MICE schemes. Together with the trace plots from Figures 4.9 to 4.11 in Section 4.7.3, this finding further supports the statement that drug-trying response

variables influenced the MICE imputation.

5.2.5 The Univariate Logistic Regression Model with the Drug-trying Response Variables and Covariates

To investigate the covariates that associate with young people's drug-trying behaviour, we expanded the univariate logistic regression models with the drug-trying response variables to include the smoking, drinking and drug-related socio-demographic covariates. These univariate logistic regression models were known as logistic regression model 2, which was stated in Section 5.2.3. In this case, only the imputed data sets from the MICE imputation scheme 2, i.e. MICE Imputation, FCS based on full data frame, were adopted for imputation of the data.

The primary goal of conducting the univariate logistic regression models with the smoking, drinking and drug-related socio-demographic covariates was to find the most parsimonious model that explained the association of the smoking, drinking and drug-trying socio-demographic covariates to each of the 15 drug-trying response variables. In other words, 15 such most parsimonious models were constructed separately each for one of the 15 corresponding drug-trying response variables. To achieve this goal with the imputed data sets, a selected set of variables in pre-defined forms (either linear or categorical (factor) variables), and the model selection by backward elimination, were required. The steps for constructing the univariate logistic regression models, i.e. model 2, were:

(1) Each potential ordinal variable was either treated as a linear variable or a categorical (factor) variable. This was because treating the variable as ordinal would lead to the univariate logistic regression models (produced from `glm()` function in `R` program) being difficult to interpret. The determination of the

variable type of these potential ordinal variables was by the Akaike Information Criterion (AIC) through complete case analysis on each variable. The resultant types of covariates for the univariate logistic regression models with covariates (i.e. model 2) are presented in Tables B.1.1 and B.1.2 in Appendix B.1.

(2) Checked if a pair of highly correlated variables could be applied to the univariate logistic regression model with a drug-trying response variable simultaneously, without any error in the GLM function in R program, a two-parameter logistic regression under complete case analysis was carried out. The two-parameter logistic regression model is described by the following equation:

$$\text{logit}(Y_i) = \beta_0 + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} \tag{5.1}$$

where for respondent $i = 1, \ldots, n$, (Y_i) represents the drug response (Y), X_{1i} and X_{2i} represent two highly correlated covariates (X_1 and X_2), whereas β_0 represents the intercept parameter of the logistic regression model, β_{1i} and β_{2i} represent the corresponding parameters of these two covariates. There were two scenarios that indicated the necessity of choosing a variable between the two covariates. The first scenario was the failure of the model fitting, which indicated the singularity between these two covariates. The second scenario was the unusual large standard errors, which indicated the high correlation between the two covariates. If either one of these two scenarios occurred, then with the same set of complete cases, two logistic regression models, involving each covariate term and its corresponding parameter in each model, were used. The two logistic regression models for two respective covariates are described by the following equations:

$$\text{Model A: logit}(Y_i) = \beta_0 + \beta_{1i}X_{1i}; \tag{5.2}$$

$$\text{Model B: logit}(Y_i) = \beta_0 + \beta_{2i}X_{2i}. \tag{5.3}$$

The AICs of both models were then calculated and compared. If the AIC of Model A was lower, then the covariate X_1 was included in the saturated models of the backward elimination by Rubin's Rule of the univariate logistic regression models (i.e. model 2) with the MICE scheme 2 imputed data sets, and X_2 was then discarded from the backward elimination, and vice versa for Model B.

These two steps were carried out separately and differently for each univariate logistic regression model with one drug-trying response variable (e.g. the two steps were carried out separately and differently in the univariate logistic regression model with Cannabis as well as in the univariate logistic regression model with cocaine). Covariates that were discarded from the initial model of each univariate logistic regression model, as well as those that were remained in the initial model, are presented in Tables B.2.1 and B.2.2 in Appendix B.2.

Another challenge of fitting the univariate logistic regression models with the drug-trying response variables was the problem of sparsity in the drug-trying response variables and the covariates, which was described in Agresti (2002). The problem could be explained by a two-by-two contingency table of two variables as shown in Table 5.2.1 below.

Table 5.2.1: Contingency Table of Two Binary Covariates, X_1 and X_2

		X_1	
		No	Yes
X_2	Level 0	n_{00}	n_{10}
	Level 1	n_{01}	$n_{11} = 0$

By referring to Table 5.2.1, the log-odds ratio of X_1 against X_2 was $(n_{00} * n_{11})/(0 * n_{10})$, which was either infinite or negative infinite. Such log-odds ratio resulted in fitting problems in statistical inference. Agresti (2002) described the problem as "empty cell" problem and offered a solution of adding 0.5 observed frequencies in the empty cell (in this example, the n_{11} cell) for improving the performance

of statistical inference of log-linear analysis. For practical application to the univariate logistic regression models, suppose X_1 was a response variable, and X_2 was a covariate variable, then we randomly selected a case which satisfied the two conditions $X_1 = 0$ and $X_2 = 1$, and converted its response of X_1 from "No" to "Yes", hence satisfying the conditions $X_1 = 1$ and $X_2 = 1$. As a result, n_{01} increased by 1, whereas n_{11} decreased by 1. These procedures preserved the total frequency of X_2 while improving the statistical inference of the univariate logistic regression models and therefore improving the credibility of the models.

Similarly, if X_4 was a categorical variable with P levels, where $P > 2$, and X_3 was a binary variable, the contingency table between X_3 and X_4 can be described as Table 5.2.2 below:

Table 5.2.2: Contingency Table of a Binary Covariate X_3 and a Multi-level Covariate X_4

		X_4				
		0	...	p	...	P
X_3	No	n_{00}	...	n_{0p}	...	n_{0P}
	Yes	n_{10}	...	$n_{1p} = 0$	...	n_{1P}

where the frequency cell that represented the conditions when $X_3 = 1$ (yes) and $X_4 = p$ was of zero count, then we randomly selected a case that satisfied both conditions $X_3 = 0$ and $X_4 = p$ and changed its response on X_3 from "No" to "Yes".

During model fitting, if the problem of singularity and high correlation still existed, then the problematic variables would be discarded one by one until no such problem remained.

5.2.5.1 Results of Logistic Regression Model within Drug-Trying Response Variables and Covariates

The results of variable selection were applied to data sets imputed under MICE scheme 2. Each cell of the covariate sign plots contained five symbols, represented combination of response and covariate, as well as under which type of data set ($**$: Positive association between a response variable and a covariate, significant at all factor levels; $*$: Positive association between a response variable and a covariate; significant not at all factor levels; *x*: Mixed association between a response variable and a covariate within factor levels; *v*: Negative association between a response variable and a covariate, significant not at all factor levels; *vv*: Negative association between a response variable and a covariate; significant at all factor levels). The tables of types of variables used in the univariate logistic regression models, as well as covariates included in the models, are shown in Appendix B.

Two sets of covariate sign plot tables, which indicated relationships and significance of the remaining covariates in the final models after backward elimination for the data sets imputed under MICE scheme 2, are displayed in Tables 5.2.3 to 5.2.6. The related tables of estimates and standard errors for the final models for the data sets, imputed under MICE scheme 2, can be referred to Tables B.4.1 to B.4.21 in Appendix B.4. The full description of each covariate can be referred to Tables A.2.1 to A.2.3 in Appendix A.

Table 5.2.3: Estimate Symbol Table in Logistic Regression (Table 1) ($**$: Positive association between a response variable and a covariate, significant at all factor levels; $*$: Positive association between a response variable and a covariate; significant not at all factor levels; *x*: Mixed association between a response variable and a covariate within factor levels; *v*: Negative association between a response variable and a covariate, significant not at all factor levels; *vv*: Negative association between a response variable and a covariate; significant at all factor levels.)

Variable	Type	Can	Her	Cok	Msh	Crk	Mth	Ecs	Amp	LSD	Pop	Ket	Ana	Gas	Oth	Trn
CgFam1	Factor					x										
CgStat1	Factor	*		*	**				*	*						
CgStat1	Linear														**	
CgPk1	Factor															
CgGet1	Factor					vv										
CgGet1	Linear															**
CgGet2	Factor			v												
CgGet2	Linear															
CgGet3	Factor															
CgGet3	Linear															
CgGet	Factor	**														
CgPp1	Factor													x		
CgWho1	Factor	*											**	vv		
CgBuyF1	Factor	x							x		**	x		x		
CgBuyF1	Linear															
CgEstim	Factor	*					x						x			
CgEstim	Linear															
CgPe1	Factor													v		
CgIn1	Factor													**		
LsSmk	Factor			*			**									

Table 5.2.4: Estimate Symbol Table in Logistic Regression (Table 2) ($**$: Positive association between a response variable and a covariate, significant at all factor levels; $*$: Positive association between a response variable and a covariate; significant not at all factor levels; *x*: Mixed association between a response variable and a covariate within factor levels; *v*: Negative association between a response variable and a covariate, significant not at all factor levels; *vv*: Negative association between a response variable and a covariate; significant at all factor levels.)

Variable	Type	Can	Her	Cok	Msh	Crk	Mth	Ecs	Amp	LSD	Pop	Ket	Ana	Gas	Oth	Trn
AlLast	Factor		x													
AlLast	Linear															
AlFreq2	Factor	**														
AlFreq2	Linear				**						**					
AlBnPub	Factor													**		**
AlEstim	Factor													v		
AlEstim	Linear															
LsAlc	Factor						vv							**		
AlPar	Factor													x		
AlPar	Linear	vv														
AlBuy1	Factor															
AlBuy1	Linear															vv
AlBuy2	Factor															
AlBuy2	Linear								*							
AlBuy	Factor															
AlUs1	Factor															
AlUs2	Factor													x		
Al4W1	Factor	**						*								

Table 5.2.5: Estimate Symbol Table in Logistic Regression (Table 3) ($**$: Positive association between a response variable and a covariate, significant at all factor levels; $*$: Positive association between a response variable and a covariate; significant not at all factor levels; *x*: Mixed association between a response variable and a covariate within factor levels; *v*: Negative association between a response variable and a covariate, significant not at all factor levels; *vv*: Negative association between a response variable and a covariate; significant at all factor levels.)

Variable	Type	Can	Her	Cok	Msh	Crk	Mth	Ecs	Amp	LSD	Pop	Ket	Ana	Gas	Oth	Trn
AlWhy1	Factor															
AlWho1	Factor													x		
AlPe1	Factor															
AlIn1	Factor															
DgPe1	Factor	x														
DgIn1	Factor					x										
DgEstim1	Factor	**							*					**		
D-gEstim1	Linear				**											
Books1	Factor													**		
Books1	Linear		vv													vv
LsDrg	Factor								vv					vv		
age	Linear	**			**					vv				vv		
gender	Factor	vv												**		
FSM1	Factor								vv						vv	
TruantN	Factor	**												**		
TruantN	Linear														**	**
ExclAN1	Factor															
ExclAN1	Linear														**	
SHA	Factor							x						x		

Table 5.2.6: Estimate Symbol Table in Logistic Regression (Table 4) ($**$: Positive association between a response variable and a covariate, significant at all factor levels; $*$: Positive association between a response variable and a covariate; significant not at all factor levels; *x*: Mixed association between a response variable and a covariate within factor levels; *v*: Negative association between a response variable and a covariate, significant not at all factor levels; *vv*: Negative association between a response variable and a covariate; significant at all factor levels.)

Variable	Type	Can	Her	Cok	Msh	Crk	Mth	Ecs	Amp	LSD	Pop	Ket	Ana	Gas	Oth	Trn
Cannabis	Factor			**	**	**	**	**		**	**	**	**			
Heroin	Factor			**		**	**					**		**	vv	
Cocaine	Factor	**	**		**	**		**	**		**					
Magic Mushrooms	Factor	**	**		**	**		**	**		**					
Crack	Factor	**	**	**											**	**
Methadone	Factor								**				**			
Ecstasy	Factor			**			**			**		**				**
Amphetamine	Factor			**	**											
LSD	Factor		**		**			**				vv	**			
Poppers	Factor	**		**						**				**		
Ketamine	Factor	**						**	**						**	
Anabolic Steroid	Factor	**					**		**	**				**	**	
Gas	Factor	**	**		**			**			**		**			
Other Drugs	Factor					**						**				**
Tranquillisers	Factor				**						**	**				

Tables 5.2.3 to 5.2.6 showed that totally there were 12 smoking covariates, 11 drinking covariates and 11 drug-related socio-demographic covariates, which were significant at 5 % significance level, in at least one of 15 univariate logistic regression models. In the following paragraphs, other important covariates that were related to trying drugs are discussed.

Firstly, the relationships between smoking behaviour and drug-trying behaviour were discussed. From the 15 univariate logistic regression models, the students who smoked more recently, more frequently and more heavily (referring to CgStat1) were more likely to try cannabis, cocaine, magic mushrooms, amphetamines, LSD or other drugs. The students who took lessons about smoking (LsSmk) were more likely to try cocaine or methadone and those who purchased cigarettes more often (CgBuyF1) were more likely to try poppers. The students who received information about smoking through people (CgPe1) were less likely to try gas. In contrast, those students who received information about smoking through media (CgIn1) were more likely to try gas. Also, the students who lived with smokers (CgWho1) were more likely to try cannabis or anabolic steroids, but they were less likely to try gas.

Secondly, the relationships between drinking behaviour and drug-trying behaviour were discussed. From the 15 univariate logistic regression models, the students who drank more recently (AlLast) were more likely to try heroin. Similarly, the students who drank more frequently (AlFreq2) were more likely to try cannabis, magic mushrooms or poppers. The students who had been in a pub, bar or club in the evening (AlBnPub) were more likely to try gas or tranquillisers, and those who had incidents after drinking in the last four weeks (Al4W1) were more likely to try cannabis or ecstasy as well. Nonetheless, the students whose drinking behaviour were tended to be supported by their families (AlPar1) were less likely to try cannabis. Those students who knew more

people of their own age (AlEstim) addicted to drinking were less likely to try gas.

Finally, we discussed the relationships between drug-related socio-demographic covariates and drug-trying behaviour. From the 15 univariate logistic regression models, the students who knew more people of their own age using drugs (DgEstim) were more likely to try cannabis, magic mushrooms, amphetamines and gas. Moreover, the students who truanted more often (TruantN) were more likely to try cannabis, gas, tranquillisers or other drugs, and those who had been more often excluded from schools (ExclAN1) were more likely to try other drugs. On the other hand, the students who possessed more books (Books1) were found to be less likely to try heroin or tranquillisers, but more likely to try gas, possibly with the reason of relieving academic stress. The older students (Age) were more likely to try cannabis or magic mushrooms, but on the other hand, less likely to try LSD or gas. Girls were more likely than boys to try gas, but less likely to try cannabis. The students whose families were less wealthy, reflected by the free school meal variable (FSM1), were less likely to try amphetamines or other drugs.

Based on the above-mentioned findings from the covariate sign plot Tables 5.2.3 to 5.2.6, it can be concluded that numerous smoking, drinking and drug-related socio-demographic covariates are associated with drug-trying behaviour in different dimensions. We defined important covariates as those which were present in at least three univariate logistic regression final models. By referring to Tables 5.2.3 to 5.2.6, these important covariates are: CgStat1, CgWho1, CgBuyF1, CgEstim, AlFreq2, DgEstim, Books1, Age and TruantN, reflecting the situation that smoking variables and drug-related socio-demographic variables are more influential than drinking variables to the student's drug-trying behaviour.

In order to further investigate how every factor and linear terms of a covariate

was related to the students' drug-trying behaviour, the estimates and standard errors of the univariate logistic regression models are presented in Tables B.4.1 to B.4.21 in Appendix B.4. These estimates and standard errors were discussed by focusing on interpreting the four univariate logistic regression final models with cannabis, gas, crack or tranquillisers as the drug-trying response variables respectively. As discussed in Section 5.2.4.1, cannabis and gas were two drugs of which a higher proportion of the students reported in the Year 2010 Survey that they had tried them, whereas crack and tranquillisers were drugs with moderate and the lowest proportion of the students who reported trying them in the Year 2010 Survey.

Firstly, the univariate logistic regression final model with cannabis as drug-trying response variable was discussed. The covariate measuring the cigarette smoking status was included in the univariate logistic regression final model. The students who had tried smoking or smoked before were the most likely group to try cannabis, at an odds ratio of $e^{2.3924} = 10.9397$. Those students were trailed by the students who smoked heavily at an odds ratio of $e^{1.5900} = 4.9037$, then by those who smoked moderately and those who smoked lightly, at odds ratios of $e^{0.8510} = 2.3420$ and $e^{0.3877} = 1.4736$ respectively. These odds ratios implied that more frequent smoking increased the likelihood of a student to try cannabis, but the most determinant factor was whether a student had smoked before and stopped smoking at the survey time. Another covariate measuring the source of obtaining cigarettes was included in the model. It was found that the students who obtained cigarettes from at least two types of sources (mixture) were most likely to try cannabis, at an odds ratio of $e^{3.1510} = 23.3594$. The students who obtained cigarettes from shops, people, or given by people, were similarly likely to try cannabis, at odds ratios of $e^{2.4625} = 11.7341$, $e^{2.8112} = 16.6299$ and $e^{2.7442} = 15.5522$ respectively. This result pinpointed that if a student obtained cigarette from more types of sources, he or she would be more likely to try

cannabis. These findings were similar to the covariate measuring the smokers in house and where the students who lived with the smokers and smoked outside or inside were more likely to try cannabis, at odds ratios of $e^{0.3740} = 1.4535$ and $e^{0.2764} = 1.3184$ respectively. On the other hand, the students who purchased cigarettes (CgBuyF1) a few times were most prone to try cannabis, at an odds ratio of $e^{0.5746} = 1.7764$. Those who purchased cigarettes occasionally were more likely to try cannabis at an odds ratio of $e^{0.5694} = 1.7672$, but those who did frequently were less likely to try cannabis, at an odds ratio of $e^{-0.8214} = 0.4398$. These findings suggested that the frequent purchase of cigarettes might suppress the behaviour of trying cannabis.

The covariate which measured the sources a student obtain information about drugs from people (DgPe1), included in the final model, revealed that the students who obtained information from parents and other relatives were more likely to try cannabis, at an odds ratio of $e^{0.4807} = 1.6172$, whereas those who obtained information from the professionals and the police were less likely to try cannabis at an odds ratio of $e^{-0.3073} = 0.7354$. For the students who obtained information from both types of sources, the dominating effect of parents and other relatives led to a slightly positive likelihood of the students to try cannabis (odds ratio: $e^{0.0875} = 1.0914$). On the other hand, the students who knew larger proportions of drug takers (DgEstim) were more likely to try cannabis (odds ratios from $e^{0.6854} = 1.9846$ to $e^{2.0144} = 7.4962$), whereas the students who were older (age) by every unit of year were more likely to try cannabis at an odds ratio of $e^{0.3752} = 1.4553$, and girls were less likely to try cannabis than boys at an odds ratio of$e^{-0.9074} = 0.4036$. Finally, a covariate that measured the frequency of truancy of the students was included in the final model. Those students who played truant a year prior of the survey were found to be more likely to try cannabis at an odds ratio of $e^{0.5708} = 1.7697$. Moreover, those students who had played truant at least three times were found to be more likely to try cannabis

at an odds ratio of $e^{0.3093} = 1.3625$.

Secondly, the univariate logistic regression final model with gas as the drug-trying response variable was discussed. A covariate recording the types of people that the students knew who smoked cigarettes (CgPp1) was included in the final model. The students who knew their friends smoking were more likely to try gas, at an odds ratio of $e^{0.4565} = 1.5785$, whereas those who knew at least two of three types of people smoking were most likely to try gas, at an odds ratio of $e^{0.561} = 1.7524$. These results reflected the influence of smoking friends on drug-trying behaviour. Moreover, another covariate recording whether the people who the students lived with smoked outside or inside their houses (CgWho1) was included in the final model. The students who lived with people smoking inside their houses were the least likely to try gas, at an odds ratio of $e^{-0.5167} = 0.5965$, whereas the students who lived with people smoking outside their houses were less likely to try gas, at a log odds ratio of $e^{-0.2887} = 0.7492$.

A covariate that recorded the frequency of purchasing cigarettes by a student (CgBuyF1) was included in the final model that the students who bought cigarettes occasionally were found to be less likely to try gas, at an odds ratio of $e^{-0.8203} = 0.4403$. Moreover, a covariate that recorded how the students obtained information about smoking from people (CgPe1) was included in the final model. The students who obtained such information from parents, relatives and professionals, police were less likely to try gas, at an odds ratio of $e^{-0.5128} = 0.5988$. Another predictor that recorded how the students obtained information about smoking from media (CgIn1) was included in the final model. In contrast to CgPe1 variable, the students who obtained such information from passive media, interactive media, or both (CgIn1), were more likely to try gas. The students who had been in a pub (AlBuPub) were also more likely to try gas, at an odds ratio of $e^{0.2807} = 1.3241$. Also, the students who knew half of

other people drinking (AlEstim) were less likely to try gas, at an odds ratio of $e^{-0.5442} = 0.5803$. The students who received lessons about drinking (LsAlc) were more likely to try gas, at an odds ratio of $e^{0.6697} = 1.9537$. On the other hand, from the variable describing the types of places a student usually uses alcohol in(AlUs2), the students who consumed alcohol at home or in a party, as well as consuming alcohol at pub, home or party, and in other places, were more likely to try gas, at odds ratios of $e^{0.6213} = 1.8613$ and $e^{0.5032} = 1.6540$ respectively.

A covariate that measured the proportion of drug-taking people a student knew (DgEstim) was included in the final model. The students who knew more than half of such people were more likely to try gas, at odds ratios of $e^{2.0303} = 7.6164$ for 'half' level and $e^{1.8536} = 6.3828$ for 'most/all' level. Another predictor that measured the number of books students possessed was also included in the final model. The students who possessed books (Books1) were more likely to try gas, at odds ratios of $e^{0.7698} = 2.1593$ and $e^{0.874} = 2.3965$ for 'few' and 'lots' levels respectively. In contrast, the students who had taken lessons about drugs were less likely to try gas, at an odds ratio of $e^{-0.3795} = 0.6720$; older students (Age) were less likely to try gas, with the likelihood in log scale decreasing at an odds ratio of $e^{-0.2997} = 0.7410$ with an increase in one year of age. Those students who played truant more seriously were found to be more likely to try gas, at an odds ratios of $e^{0.55} = 1.7333$, $e^{0.5799} = 1.7859$ and then $e^{0.757} = 2.1319$ if a student played truant a year ago, had played truant once or twice in the last year, and at least three times in the last year respectively. Finally, the students living in London SHA region (SHA) were more likely to try gas at an odds ratio of $e^{0.5488} = 1.7312$.

On the other hand, another two univariate logistic regression final models with crack as well as tranquillisers as drug-trying response variables were conducted respectively. For the final model of crack, firstly, the variable that predicted the family's attitude towards smoking (CgFam1) was included in the final model.

If the students' families supported the student's smoking behaviour, those students were more likely to try crack at an odds ratio of $e^{1.6797} = 5.3639$. In contrast, a covariate which measured the number of sources the students purchased cigarettes usually through shops/machine/Internet (CgGet1) was included in the final model. The students who purchased cigarettes from more than one source were less likely than those who purchased cigarettes from only one source to try crack, when compared respective odds ratios of $e^{-2.2321} = 0.1073$ and $e^{-1.3446} = 0.2606$ respectively. Moreover, the students who obtained information about drugs through passive media or through both passive and interactive media were found to be less likely to try crack, at odds ratios of $e^{-2.2219} = 0.1084$ and $e^{-1.2256} = 0.2936$ respectively.

For the univariate logistic regression final model with tranquillisers as drug-trying response variable, firstly, the predictor which measured the number of sources the students purchased cigarettes usually through shops/ machine/ Internet (CgGet1) was included in final model. The students purchased cigarettes in this way from more sources were more likely to try tranquillisers, at an odds ratio increasing by a factor of $e^{1.3055} = 3.6895$ for an increase in every level of CgGet1 variable. Besides, those students who had been in a pub (AlBnPub) were more likely to try tranquillisers at an odds ratio of $e^{0.9311} = 2.5373$. In contrast, the students who purchased alcohols from shops (AlBuy1) from at least one source were less likely to try tranquillisers at an odds ratio of $e^{-1.1832} = 0.3063$ for an increase in every level of (AlBuy1) variable; the students who possessed more books were less likely to try tranquillisers, at an odds ratio decreasing by a factor of $e^{-0.8351} = 0.4338$ for an increase in every level of Books1 variable. Finally, the students who had truanted were more likely to try tranquillisers, at an odds ratio of $e^{0.4045} = 1.4986$ for an increase in every level of the Truant variable.

To determine in the univariate logistic regression final model with a drug-trying response variable, whether the terms of drug covariates were replaced by the terms of the smoking, drinking and drug-related socio-demographic covariates, the final models containing only drug covariate terms were compared with the final models containing drug covariate terms as well as the smoking, drinking and drug-related socio-demographic terms. In that regard, only the comparable final models in respect of cannabis, gas, crack and tranquillisers were discussed.

In the final model of cannabis including the smoking, drinking and drug-related socio-demographic covariates, methadone, ecstasy and amphetamines covariates were explained by a plenty of the smoking, drinking and drug-related socio-demographic covariate terms as mentioned in previous paragraphs. The common terms in the final model yielded apparently different estimates and standard errors. On the other hand, in the final model of gas including the smoking, drinking and drug-related socio-demographic covariates, the cannabis and ecstasy terms were explained by numerous smoking, drinking and drug-related socio-demographic covariate terms, but the estimates and standard errors of the common terms in the final model were similar with those in the final model containing only drug covariates.

In the final model of crack, the other drugs term was explained by CgFam1, CgGet1 and DgIn1 variables, but the estimates and standard errors of the common terms were similar with those in the final model containing only drug covariates. Similarly, in the final model of tranquillisers including the smoking, drinking and drug-related socio-demographic covariates, the cannabis and ketamine covariate terms were explained by the ecstasy term, CgGet1, AlBnPub, AlBuy1, Books1 and TruantN predictors, but the estimates and standard errors of the common terms were quite similar with those in the final model containing only drug covariates.

5.2.6 Summary of Main Findings from Univariate Logistic Regression Analysis

In both Sections 5.2.4 and 5.2.5, univariate logistic regression analysis was employed to further study the relationships among drug-trying response variables and the smoking, drinking and drug-related socio-demographic covariates. When only involved with the 15 drug-trying response variables, univariate logistic regression analysis revealed that almost every drug has a positive interaction with other drugs (except the relationship between LSD and ketamine) but the extent of association varied among the 15 drugs. For example, cannabis was found to have positive interactions with 10 drugs (MICE scheme 2), whereas for methadone, crack and other drugs each has positive interaction with 4 drugs (MICE scheme 2). This finding indicates that using other types of drugs by a student is a good predictor of whether the student uses cannabis or not.

When including the smoking, drinking and drug-related socio-demographic covariates in the univariate logistic regression analysis, it was found that numerous smoking, drinking and drug-related socio-demographic covariates were associated with drug-trying behaviour in different dimensions. Among these smoking, drinking and drug-related socio-demographic covariates, there were important covariates which were associated with at least three of the 15 drugs. These important covariates included: (1) cigarette smoking status of a student (CgStat1); (2) number of smokers in a student's house and where they smoked (CgWho1); (3) frequency of buying cigarettes from shop by a student (CgBuyF1); (4) how many peer smokers a student knew (CgEstim); (5) usual frequency of drinking alcohol by a student (AlFreq2); (6) number of peer drug users a student knew (DgEstim); (7) how many books in a student's home (Books1); (8) age of a student (Age) and (9) how often a student played truant (TruantN).

5.3 Log-linear Analysis Model

5.3.1 Introduction

As mentioned in Vermunt (1996), a log-linear analysis model is widely used for analysing frequency tables and contingency tables. A log-linear analysis model is used to analyse the multivariate frequency tables with a set of parameters (Vermunt, 1997). A Poisson link function is employed for modelling such observed frequencies or counts.

Log-linear analysis models are also applied in behavioural studies, where case frequencies within a certain period are usually recorded (McCullagh and Nelder, 1999).

In this research, in addition to the univariate logistic regression models, a log-linear analysis model is adopted to analyse the two-way interactions among the 15 drug-trying response variables. However, we did not include three or more ways of interaction terms, since there were too few cases with three or more ways of interaction terms for modelling in R program. In the univariate logistic regression models, only one-way interaction among the drug-trying response variables can be modelled in a single regression model. In a log-linear analysis model, we can fit the patterns among the 15 drug-trying response variables with hierarchical two-way interactions, in order to investigate the relationships between these drug-trying response variables in both directions in a single model. We can also include intercepts in the log-linear analysis model to measure the probability of trying each drug by the students. According to Christensen (1997), the advantages of using log-linear analysis models are: (1) log-linear analysis models possess the properties of modelling flexibility that are associated with ANOVA and regression and (2) log-linear analysis models are easily

interpretable in terms of odd and independence. However, the disadvantage of a log-linear analysis model is that it focuses on aggregate data level rather than individual data level, as the data set for a log-linear analysis model records the frequency of each data pattern at aggregate level (Bijleveld et al., 1998).

5.3.2 Theory

In this research, the associations among 15 drug-trying response variables were evaluated by hierarchical log-linear analysis models that contained two-way interactions among drug-trying response variables. Since a hierarchical log-linear model was adopted in our analysis, only two-way hierarchical log-linear analysis model was discussed in this section.

Vermunt (1997) defined a saturated two-way hierarchical log-linear analysis model as follows: suppose there exists a frequency table with three binary variables, denoted as A, B and C. Let a, b and c be indices associated with A, B and C respectively. Let μ_{abc} be the expected frequency for the cell that belongs to category a of A, b of B, and c of C, then the equation of a saturated two-way hierarchical log-linear analysis model is expressed as Equation 5.4.

$$\mu_{abc} = u + u_a^A + u_b^B + u_c^C + u_{ab}^{AB} + u_{ac}^{AC} + u_{bc}^{BC}, \tag{5.4}$$

(Vermunt, 1997)

where u_a^A, u_b^B and u_c^C indicate the relative number of cases at the various levels of A, B and C, and u_{ab}^{AB}, u_{ac}^{AC} and u_{bc}^{BC} represent the strength of the partial associations between A and B, A and C and B and C respectively.

The saturated two-way P-dimensional log-linear analysis model includes all the possible intercept terms and two-way interaction terms of P variables. The total number of the two-way interaction terms is $(P-1)(P-2)/2$.

To explain the log-linear equation adopted in this research, we adopt the index i associated with a pattern that contained a distinct combination of (a,b,c). A data set containing three binary variables, namely A, B and C, is adopted. Corresponding binary responses of A, B and C, are recorded for each of the respondent i, $i = 1,\ldots,n$. The data patterns for these three binary variables are illustrated in the following frequency table of all combinations of patterns in the following Table 5.3.1.

Table 5.3.1: Pattern Table of Data Set with Three Variables, A_h, B_h and C_h

h	A_h	B_h	C_h	Frequency
1	0	0	0	F_{000}
2	0	0	1	F_{001}
3	0	1	0	F_{010}
4	0	1	1	F_{011}
5	1	0	0	F_{100}
6	1	0	1	F_{101}
7	1	1	0	F_{110}
8	1	1	1	F_{111}

The equation of the corresponding log-linear analysis model, containing one and two-dimensional interactions, is expressed as the following:

$$\mu_h = u + u_h^A + u_h^B + u_h^C + u_h^{AB} + u_h^{AC} + u_h^{BC}, \tag{5.5}$$

where μ_h is the expected frequency for pattern h, $h = 1,\ldots,8$, u is the global intercept parameter and u_h^A, u_h^B, u_h^C are main effects for binary variables A,B,C, and $u_h^{AB}, u_h^{AC}, u_h^{BC}$ are parameters representing interactions between A and B, A and C and B and C, respectively.

An alternative log-linear model formulation for the frequency of each pattern h,

denoted as F_{A_h,B_h,C_h}, of A_h, B_h, C_h, is expressed as:

$$\log(\mathbf{E}[F_{A_h,B_h,C_h}]) = \lambda_0 + \lambda_A A_h + \lambda_B B_h + \lambda_C C_h + \lambda_{AB} A_h B_h + \lambda_{AC} A_h C_h + \lambda_{BC} B_h C_h, \quad (5.6)$$

where $\mathbf{E}[F_{A_h,B_h,C_h}]$ is the expected frequency for pattern h, $h = 1,\ldots,8$, for observed values of A_h, B_h, C_h, λ_0 is the global intercept parameter (for the zero vector pattern), and $\lambda_A, \lambda_B, \lambda_C$ are effects associated with A, B, C, and $\lambda_{AB}, \lambda_{AC}, \lambda_{BC}$ are parameters for two-way interaction terms.

When conducting a log-linear analysis model (where the combinations are modelled by a Poisson GLM) to pattern data, denoted $X'' = \{x''_1,\ldots,x''_h,\ldots,x''_8\}$, where x''_h represents the h^{th} data pattern, with associated response vector (frequency vector) $Y'' = \{y''_1,\ldots,y''_h,\ldots,y''_8\}$, an appropriate link function is the $\log(\mu_h)$ link, where the μ_h is the expected frequency for pattern h. We have:

$$y''_h \sim \text{Poisson}(\mu_h), \quad (5.7)$$

$$\mathbf{E}\left[y''_h\right] = \mu_h, \quad (5.8)$$

$$\log(\mu_h) = \beta_0 + (x''_h)^T \beta. \quad (5.9)$$

where β is a vector representing main effects and interaction terms, and β_0 is an intercept of the model. The log-likelihood for the log-linear analysis model is expressed by the following equation:

$$\ell(\mu, Y'') = \sum_{h=1}^{8} (y''_h \log(\mu_h) - \mu_h - \log(y''_h!)).$$

Derived from Equation 5.5, the expected log frequencies in this pattern table

are presented by the following equations.

$$\log(\mathbf{E}[F_{000}]) = \lambda_0; \tag{5.10}$$

$$\log(\mathbf{E}[F_{010}]) = \lambda_0 + \lambda_B; \tag{5.11}$$

$$\log(\mathbf{E}[F_{001}]) = \lambda_0 + \lambda_C; \tag{5.12}$$

$$\log(\mathbf{E}[F_{011}]) = \lambda_0 + \lambda_B + \lambda_C + \lambda_{BC}; \tag{5.13}$$

$$\log(\mathbf{E}[F_{100}]) = \lambda_0 + \lambda_A; \tag{5.14}$$

$$\log(\mathbf{E}[F_{110}]) = \lambda_0 + \lambda_A + \lambda_B + \lambda_{AB}; \tag{5.15}$$

$$\log(\mathbf{E}[F_{101}]) = \lambda_0 + \lambda_A + \lambda_C + \lambda_{AC}; \tag{5.16}$$

$$\log(\mathbf{E}[F_{111}]) = \lambda_0 + \lambda_A + \lambda_B + \lambda_C + \lambda_{AB} + \lambda_{AC} + \lambda_{BC}. \tag{5.17}$$

Note that using the notation in Equation 5.9 above, we can hence write: $\beta_0 = \lambda_0, \beta = (\lambda_A, \lambda_B, \lambda_C, \lambda_{AB}, \lambda_{AC}, \lambda_{BC})$. By combining Equations 5.10 to 5.17, the intercepts and interaction term parameters of the log-linear analysis model are defined as follows:

$$\log(\mathbf{E}[F_{100}]) - \log(\mathbf{E}[F_{000}]) = \log\left(\frac{\mathbf{E}[F_{100}]}{\mathbf{E}[F_{000}]}\right) = \lambda_A, \tag{5.18}$$

$$\log(\mathbf{E}[F_{010}]) - \log(\mathbf{E}[F_{000}]) = \log\left(\frac{\mathbf{E}[F_{010}]}{\mathbf{E}[F_{000}]}\right) = \lambda_B, \tag{5.19}$$

$$\log(\mathbf{E}[F_{001}]) - \log(\mathbf{E}[F_{000}]) = \log\left(\frac{\mathbf{E}[F_{001}]}{\mathbf{E}[F_{000}]}\right) = \lambda_C, \tag{5.20}$$

$$\log(\mathbf{E}[F_{110}]) - \log(\mathbf{E}[F_{100}]) - [\log(\mathbf{E}[F_{010}]) - \log(\mathbf{E}[F_{000}])] = \lambda_{AB}, \tag{5.21}$$

$$\log(\mathbf{E}[F_{101}]) - \log(\mathbf{E}[F_{100}]) - [\log(\mathbf{E}[F_{001}]) - \log(\mathbf{E}[F_{000}])] = \lambda_{AC}, \tag{5.22}$$

$$\log(\mathbf{E}[F_{011}]) - \log(\mathbf{E}[F_{010}]) - [\log(\mathbf{E}[F_{001}]) - \log(\mathbf{E}[F_{000}])] = \lambda_{BC}. \tag{5.23}$$

From Equation 5.10, λ_0 represents the expected log frequency of respondents at the baseline. The derivation of intercept parameters, $\lambda_A, \lambda_B, \lambda_C$ in Equations 5.18 to 5.20, shows that these parameters represent the log odds of variables A, B and

C respectively, given the condition of zero as rest of responses for a respondent. Finally, the derivation of the respective interaction parameters, $\lambda_{AB}, \lambda_{AC}, \lambda_{BC}$, implies these parameters represent log odds ratio between any pair of variables, given the condition of zero as rest of responses for a respondent.

In general, the intercept term of the log-linear analysis model for variable X is the log-odds of the corresponding variable X, whereas the interaction term for variable X and Y is the log-odds ratio of the two corresponding variables X and Y.

Same as the logistic regression analysis models, Rubin's rule with Backward Elimination with Wald's test can be applied to log-linear analysis. Rubin's rule and Wald's test can be referred to Section 4.5.3.3 respectively. Backward Elimination begins with the inclusion of all relevant intercept terms and interaction terms in a Poisson Generalized Linear Model, known as the saturated model. Such model is fitted to all M imputed data sets. Estimates and standard errors from the M imputed data sets are combined and pooled by Rubin's Rule. Wald's test is then conducted for each of the estimates in the model. If the combined p-value of an estimate is higher than 0.05, then the term is discarded from the model; if not, then the variable is retained in the model. The term with the highest combined p-value (here, we denote it as X_{P1}) is discarded from the model at each step. The M imputed data sets are re-fitted without the X_{P1} term, and the process repeats, where the term with the highest combined p-value, X_{P2} at this step, is discarded. The process repeats until no insignificant terms remain in the model. Such model at this status is considered as the final model.

For complete case analysis, Wald's test was adopted as the term selection test for the backward elimination.

The next step is extending the log-linear analysis model to include four response variables, A, B, C and D, up to two-way interactions. Similar with the log-linear analysis model with three response variables, the equation for two-way log-linear analysis model with four response variables is expressed below:

$$\mu_h = u + u_h^A + u_h^B + u_h^C + u_h^D + u_h^{AB} + u_h^{AC} + u_h^{AD} + u_h^{BC} + u_h^{BD} + u_h^{CD}. \tag{5.24}$$

The two-way log-linear analysis model can be extended further to include P response variables.

5.3.3 Application of Log-linear Analysis Model

In this research, a log-linear analysis model was adopted with the objective of further investigating the relationships between 15 drug-trying response variables. The log-linear analysis model was fitted for data sets of two imputation schemes and complete case analysis. The two imputation schemes were:

Scheme 1: MICE Imputation, FCS based upon 15 drug-trying response variables only;
Scheme 2: MICE Imputation, FCS based upon full data set;

In this research, the two-way interactions between 15 drug-trying response variables were investigated. The most parsimonious model was obtained through the backward elimination. Since the log-linear analysis model was a hierarchical model, intercept terms were required.

Similar to the univariate logistic regression model, one selection process for a log-linear analysis model was to carry out the backward elimination. The backward elimination for the log-linear analysis model commenced from the saturated model that contained all the intercepts and two-way interaction terms

only from the ten pattern tables of ten respective imputed data sets imputed by each imputation scheme (scheme 1 and scheme 2). Details of backward elimination could be found in Section 5.2.2. Any pattern with zero predicted frequency was discarded from the pattern tables before conducting the log-linear analysis.

In the following sections, firstly we discuss the results of the final model of Log-linear Analysis with backward elimination, in Section 5.3.4.1. Then we discuss the results of saturated models of log-linear analysis, in Section 5.3.4.2. Both result and discussion sections begin with a log-odds ratio heat plot of the resultant model, as well as a covariate sign plot, which describes whether an interaction term between any combination of two drug variables, under any scheme, is positive (indicated in blue), negative (indicated in red), or non-significant at 5 % significance level (indicated in grey). Discussions about these plots are then followed, and the whole section ends with a conclusion.

5.3.4 Results and Discussion

5.3.4.1 Results of Final Log-linear Analysis Model with Backward Elimination

In this section, the tables of estimates of log-linear analysis models, as well as the covariate sign plot, are discussed, with emphasis on which pairs of drug-trying response variables existed in the final models and their relationships. The tables of estimates of the final log-linear analysis models are presented in Tables C.1.1 to C.1.3 in Appendix C, whereas those of saturated log-linear analysis models are presented in Tables C.2.1 to C.2.3. The log-odds ratio heat plots of the final log-linear analysis models for two MICE schemes are shown as Figure 5.5, while the model's covariate sign plot is shown as Figure 5.6.

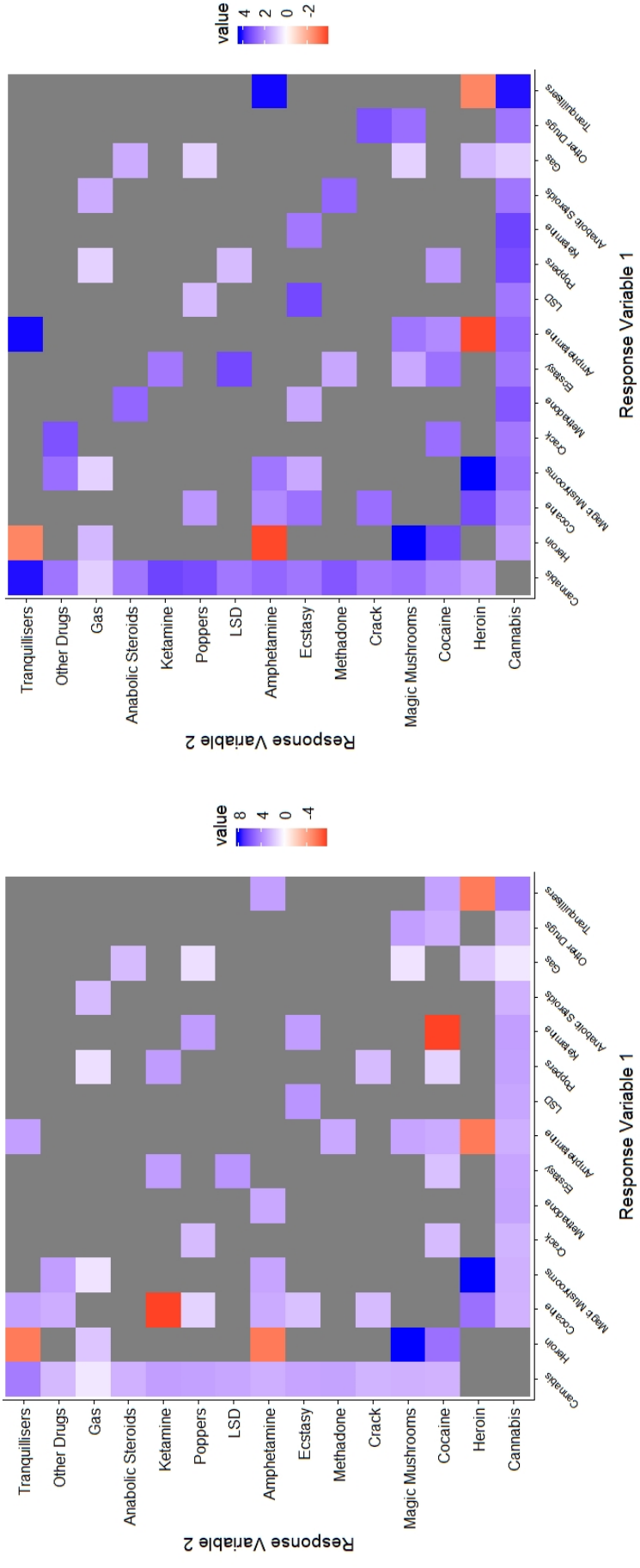

Figure 5.5: Log-odds Ratio Heat Plot of Log-linear Analysis Final Model under Scheme 1 (left) and Scheme 2 (right). Blue represents positive log-odds ratio, red represents negative log-odds ratio and grey represents insignificant or non-applicable log-odds ratios.

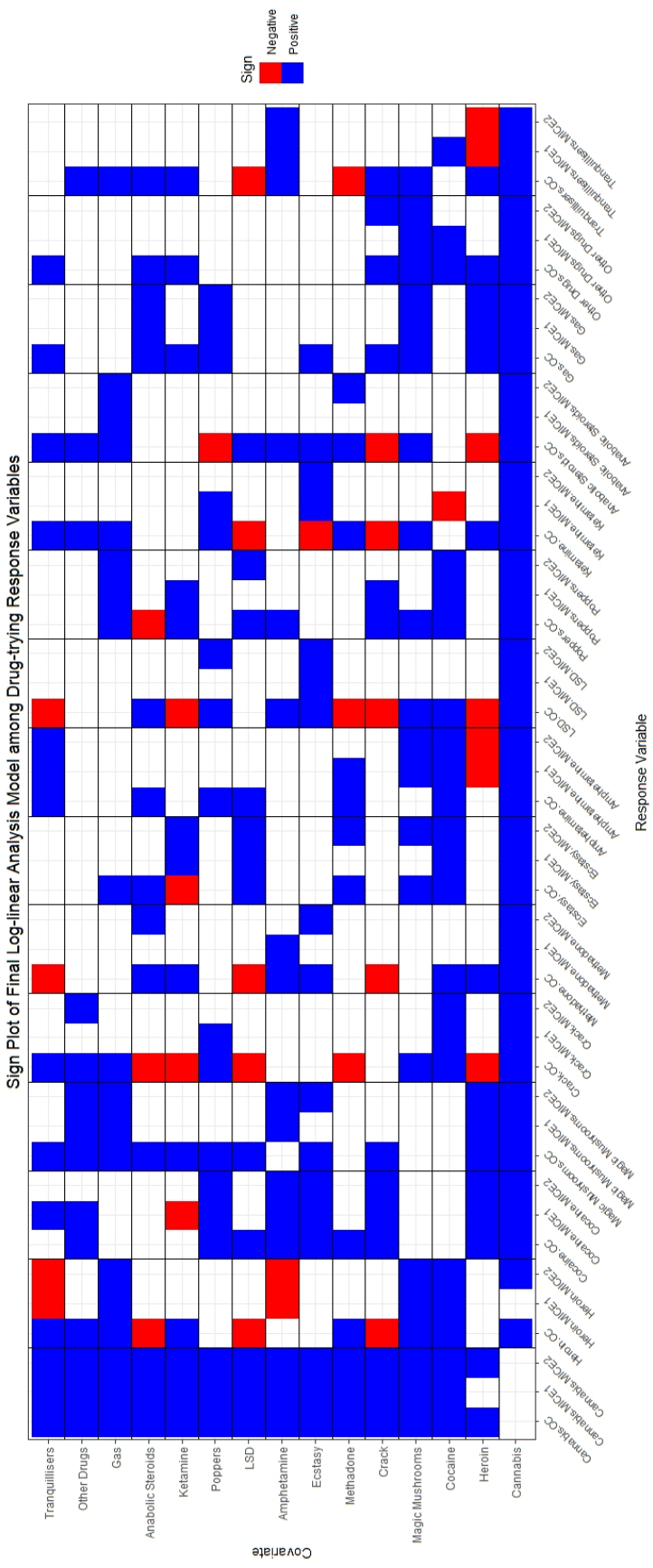

Figure 5.6: Covariate Sign Plot of Log-linear Analysis Final Model (Axis label description - CC: Complete case analysis; MICE1: MICE Imputation Scheme 1; MICE 2: MICE Imputation Scheme 2) (Grid Colour Description - Blue: Positive Association; Red: Negative Association; White: No Significant Association at 5% significance level, Not applicable)

When comparing the estimates and standard errors of the interaction terms presented in the final log-linear analysis models, as shown in Tables C.1.1 to C.1.3, for both imputation schemes 1 and 2, it was observed that many estimates were similar. Some interaction terms with large absolute values were different between the results of the two imputation schemes, for example, the interaction term for heroin and tranquillisers. However, signs of estimates of all the interaction terms in respect of two imputation schemes matched with each other and most standard error values of the interaction terms were quite similar. This observation is reflected in Figure 5.5. When we compared the estimates and standard errors of the interaction terms presented in the final models between the complete case analysis and the two MICE imputation schemes, the interaction terms presented in a pair of final models for two MICE imputation schemes were substantially different from those for the complete case analysis, especially the interaction terms that were related to heroin, cocaine, magic mushrooms, crack, methadone, ecstasy, amphetamine, LSD and poppers. These differences could be explained that drug-trying response variables influenced the MICE imputation, thus producing different results from the complete case analysis.

When comparing the model results from the different imputation schemes, we observed that these results appeared to be slightly different, suggesting the influence of including covariates other than drug-trying response variables in multiple imputation.

When comparing the results of final log-linear analysis models based upon complete case analysis with those of log-linear analysis models based upon two MICE imputation schemes in Figure 5.6, an interaction term, i.e. the interaction term between heroin and tranquillisers, has conflicting directions. This discrepancy might be caused by the adjustments from other interaction terms in the two log-linear analysis models due to differences in imputed missing data.

From Figure 5.5, the first observation was that cannabis yielded the largest number of significant interactions with other types of drugs. The two imputation schemes generally agreed with most of the significant covariates between cannabis, heroin, magic mushrooms, ecstasy, amphetamines, LSD, poppers and anabolic steroids predictors. When investigating the significant interaction terms in final log-linear analysis models based upon MICE scheme 2, the results of the final log-linear analysis models largely agreed with the univariate logistic regression final models with various drug-trying response variables as the responses. The model with heroin, cocaine, magic mushrooms or ecstasy as the response variable yielded six other types of drug exploratory variables, indicating that though the percentages of students trying heroin, cocaine, magic mushrooms and ecstasy were tiny (0.49%, 1.19 %, 1.49% and 1.10%), heroin, cocaine, magic mushrooms and ecstasy were essential in connecting other drug-trying response variables. Additionally, models with amphetamines or gas as the response variable yielded five other types of drug exploratory variables, whereas a model with popper as the response variable yielded four other types of drug exploratory variables. Models with crack, methadone, LSD or anabolic steroids as the response variable yielded three other types of drug exploratory variables, and models with ketamine as the response variable yielded two other types of drug exploratory variables.

From Tables C.1.1 to C.1.3, the intercept terms of each of the log-linear analysis models showed that the ordering of the drug-trying response variables, in terms of the proportion of the students trying them, were generally similar with the corresponding ordering of the students trying each of the 15 drugs as shown in Table 3.1.6. For example, the log-odds estimates of trying cannabis and trying gas showed that they were the two drugs with higher proportions of the students trying them, which corresponded to the finding of the higher

proportions of the students trying these two drugs as shown in Table 3.1.6. The slight discrepancies existed in some drugs between the above mentioned two orderings were due to adjustments made under different imputation schemes.

Most of the log-odds ratios of the interaction terms were positive, as depicted in Figure 5.5, indicating students trying one drug were more likely to try another drug. Specifically, these positive interactions briefly explained the ordering of the proportion of students trying each of the 15 drugs. The smaller the proportion of students trying one of two drugs in a pair of interaction terms, the greater the absolute estimate value of the corresponding log-odds ratio.

The results of the MICE scheme 1 were chosen to discuss the results of the final log-linear analysis models with backward elimination, since the MICE scheme 1 considered only 15 drugs for imputation, which is in line with log-linear analysis models with drugs only. When looking at the results of the MICE scheme 1, several interaction terms with distinctive estimates were observed. The interaction term between heroin and magic mushrooms yielded the highest estimate of a log-odds ratio (8.0758), indicating each student having tried heroin was almost certain to try magic mushrooms or vice versa. On the other hand, the interaction term between cocaine and ketamine yielded the lowest estimate (-7.1597), indicating each student having tried cocaine was almost certain not to try ketamine or vice versa. Other distinctive interaction terms with positive associations include heroin and amphetamines as well as cocaine and poppers, which all highlighted the positive effects of including these interaction terms in frequencies related to these terms. Distinctive interaction terms with negative associations include heroin and amphetamines as well as heroin and tranquillisers, which all highlighted their negative effects in related frequencies.

In Section 5.3.4.2 below, the results of saturated log-linear analysis models are

discussed in a similar format.

5.3.4.2 Results of Saturated Log-linear Analysis Model

In this section, we discussed the tables of estimates and standard errors of saturated log-linear analysis models, as well as the related sign plots, with emphasis on which pairs of drug responses existed in the saturated model and their relationships. The tables of estimates are presented in Tables C.2.1 to Tables C.2.3 Appendix C. The log-odds ratio heat plots of the saturated log-linear analysis models for two MICE schemes are shown as Figure 5.7, while the models' covariate sign plot is shown as Figure 5.8.

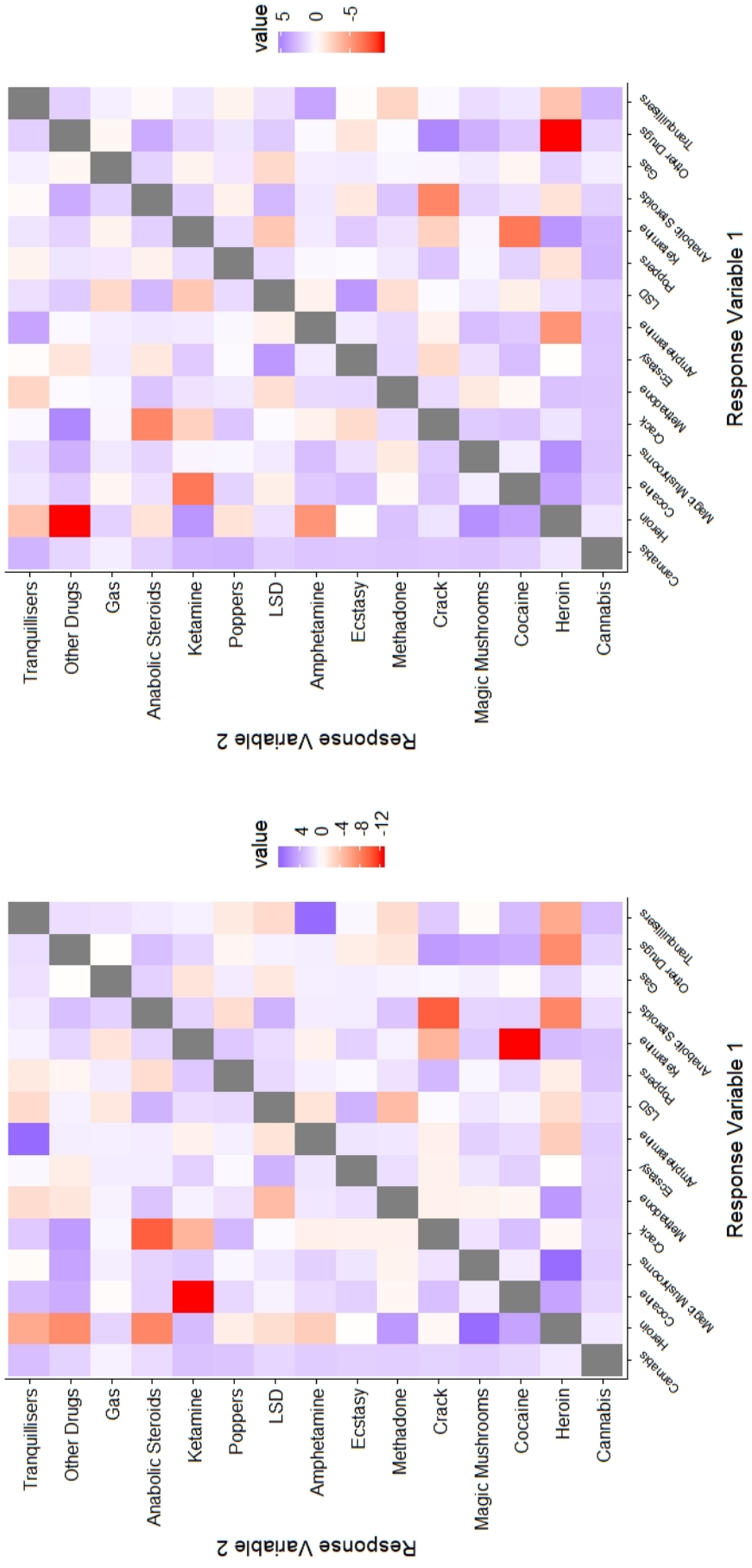

Figure 5.7: Log-odds Ratio Heat Plot of Log-linear Analysis Saturated Model under Scheme 1 (left) and Scheme 2 (right). Blue represents positive log-odds ratio, red represents negative log-odds ratio and grey represents insignificant or non-applicable log-odds ratios.

Judging from the log-odds ratio heat plots in Figure 5.7, cannabis was related to all other types of drugs except heroin, due to all mildly positive interaction terms between the cannabis and all other drug-trying response variables. Another point was that more than half of the interaction terms in the saturated models were positive, but several very negative interaction terms were found. For example, heroin was negatively correlated with other drugs and tranquillisers. On the contrary, heroin was positively correlated with methadone. This correlation was sensible because methadone was a derivation of heroin. Finally, most correlations in the saturated model were weak. One instance was that gas was weakly correlated with all other types of drugs with reference to Figure 5.7.

Figure 5.8 was considered by comparing the significant interaction terms presented in complete case analysis with those of MICE scheme 1 and MICE scheme 2 models.

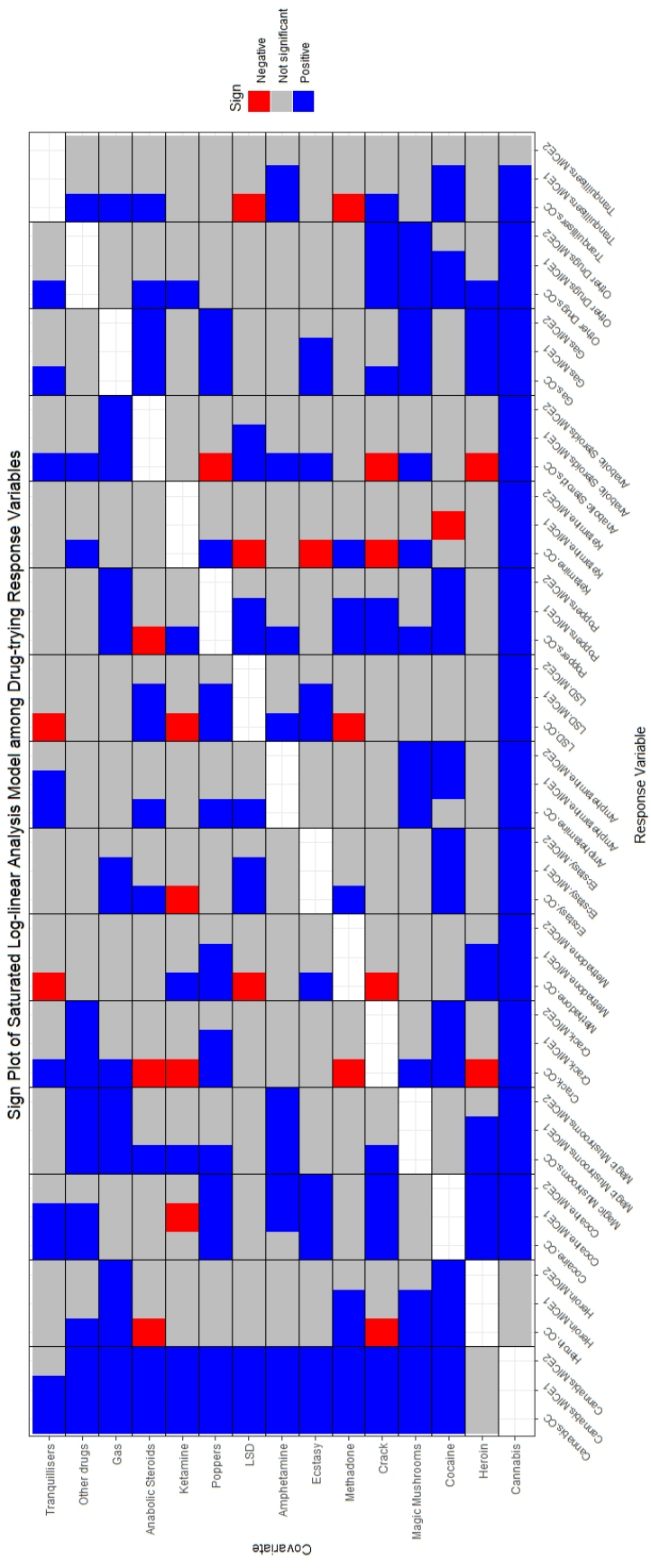

Figure 5.8: Covariate Sign Plot of Log-linear Analysis Saturated Model (Axis label description - CC: Complete case analysis; MICE1: MICE Imputation Scheme 1; MICE 2: MICE Imputation Scheme 2) (Grid Colour Description - Blue: Positive Association; Red: Negative Association; Grey: No Significant Association at 5% significance level; White: Not applicable)

Referring to Figure 5.8, concentrating on significant interaction terms only, there were no conflicting signs between complete case analysis and the two imputation schemes. Similar with the final models with backward elimination, most of the log-odds ratios of the significant interaction terms were positive, as depicted in Figure 5.8.

For saturated log-linear analysis models, estimate tables in Appendix B.3 were considered by comparing the estimates and standard errors between MICE imputation scheme 1 and MICE imputation scheme 2. From Tables C.2.1 to C.2.3, most interaction term estimates were observed to be similar. Only a few interaction term estimates, such as crack and ketamine, amphetamines and other drugs, heroin and tranquillisers, yielded relatively large differences.

From Tables C.2.1 to C.2.3, focusing on results from MICE imputation Scheme 1, several interaction terms with distinctive estimates were observed. For example, the interaction term between amphetamines and tranquillisers yielded the highest estimate (7.9116), indicating that the estimated number of students who have tried amphetamines were almost certain to try tranquillisers or vice versa. On the other hand, the interaction term between cocaine and ketamine yielded the lowest estimate (-12.3839), indicating that less estimated number of students who have tried cocaine were almost certain not to try ketamine or vice versa. Other distinctive interaction terms with positive associations include cannabis and ketamine, heroin and methadone, heroin and magic mushrooms, cocaine and other drugs. On the other hand, distinctive interaction terms with negative associations include heroin and other drugs, heroin and tranquillisers, crack and ketamine, which all highlighted their negative effects on related frequencies.

5.3.5 Comparison of Log-linear Analysis Model with Univariate Logistic Regression Models with Drug Covariates Only

Another aim of conducting the log-linear analysis models is to justify the findings of the association of a student trying a drug and another drug through comparing corresponding findings of univariate logistic regression models. In this comparison, the interaction terms of the two-way relationships between drug-trying response variables in the log-linear analysis models were compared with the corresponding interaction terms of the drug predictors to drug-trying response variables in univariate logistic regression models, which were regarded as one-way relationships among drug-trying response variables. A comparison was conducted to identify common terms and the direction of these common terms.

In respect of association and interaction among the 15 drug-trying response variables, results from the univariate logistic regression models were generally found to be comparable with those from the log-linear analysis models.

To explain this comparison, a data set containing two drug-trying response variables and n students was used. Suppose there are two binary drug-trying response variables, A and B, where the response 0 denotes "No", and another response 1 denotes "Yes", then the contingency table is presented as follows: This

		B	
		Yes	No
A	Yes	F_{11}	F_{10}
	No	F_{01}	F_{00}

contingency table can be transformed into the pattern table and the log-linear analysis model, in this case, is defined as below:

h	A_h	B_h	Frequency
1	0	0	F_{00}
2	0	1	F_{01}
3	1	0	F_{10}
4	1	1	F_{11}

$$\log(F_{A_h,B_h}) = \lambda_0 + \lambda_A A_h + \lambda_B B_h + \lambda_{AB} A_h B_h, \tag{5.25}$$

where $A_h = 0, 1$ and $B_h = 0, 1$ respectively. As a result, the prediction for all frequencies in this contingency table is listed as below:

$$\log(\mathbf{E}[F_{00}]) = \lambda_0; \tag{5.26}$$

$$\log(\mathbf{E}[F_{10}]) = \lambda_0 + \lambda_A; \tag{5.27}$$

$$\log(\mathbf{E}[F_{01}]) = \lambda_0 + \lambda_B; \tag{5.28}$$

$$\log(\mathbf{E}[F_{11}]) = \lambda_0 + \lambda_A + \lambda_B + \lambda_{AB}. \tag{5.29}$$

Referring back to the working data sets, for respondent $i = 1, \ldots, n$, the logistic regression model, in this case, is defined as below:

$$\begin{aligned}
\text{logit}(p(A_i = 1 \mid B_i)) &= \log\left(\frac{p(A_i = 1 \mid B_i)}{p(A_i = 0 \mid B_i)}\right) \\
&= \beta_{01} + \beta_{11} B_i, \\
\text{logit}(p(B_i = 1 \mid A_i)) &= \log\left(\frac{p(B_i = 1 \mid A_i)}{p(B_i = 0 \mid A_i)}\right) \\
&= \beta_{02} + \beta_{12} A_i.
\end{aligned}$$

The predicted probability $p(A_i = a \mid B_i = b)$ can be interpreted as $\mathbf{E}[F_{ab}]/\mathbf{E}[F_{B_i=b}]$, where $F_{B_i=b}$ is the frequency of cases fulfilling the condition $B_i = b$, since this is a conditional probability that $A_i = a$ given a condition $B_i = b$. Hence, in a case

when $B_i = 1$,

$$\log\left(\frac{p(A_i = 1 \mid B_i = 1)}{p(A_i = 0 \mid B_i = 1)}\right) = \log\left(\frac{(\mathbf{E}[F_{11}]/\mathbf{E}[F_{B_i=1}])}{(\mathbf{E}[F_{01}]/\mathbf{E}[F_{B_i=1}])}\right). \tag{5.30}$$

$$= \log\left(\frac{\mathbf{E}[F_{11}]}{\mathbf{E}[F_{01}]}\right) = \lambda_A + \lambda_{AB}. \tag{5.31}$$

In another case, when $B_i = 0$,

$$\log\left(\frac{p(A_i = 1 \mid B_i = 0)}{p(A_i = 0 \mid B_i = 0)}\right) = \log\left(\frac{(\mathbf{E}[F_{10}]/\mathbf{E}[F_{B_i=0}])}{(\mathbf{E}[F_{00}]/\mathbf{E}[F_{B_i=0}])}\right) = \log\left(\frac{\mathbf{E}[F_{10}]}{\mathbf{E}[F_{00}]}\right) = \lambda_A. \tag{5.32}$$

As a result, the result for λ_{AB} is as follows:

$$(5.31) - (5.32), \log\left(\frac{\mathbf{E}[F_{11}]\mathbf{E}[F_{00}]}{\mathbf{E}[F_{01}]\mathbf{E}[F_{10}]}\right) = \lambda_{AB}. \tag{5.33}$$

Indicating the interaction term in the log-linear analysis model, λ_{AB} is the log-odds ratio between A and B. The Equation 5.32 reveals λ_A in the log-linear analysis model is the log-odds for A given the condition of $B_i = 0$. Similarly, λ_B in the log-linear analysis model is the log odds for B given the condition of $A_i = 0$.

When deriving the logistic regression model, in the condition of $B_i = 0$,

$$\text{logit}(p(A_i = 1 \mid B_i = 0)) = \log\left(\frac{p(A_i = 1 \mid B_i = 0)}{p(A_i = 0 \mid B_i = 0)}\right) = \beta_{01} = \lambda_A$$

$$= \log\left(\frac{\mathbf{E}[F_{10}]}{\mathbf{E}[F_{00}]}\right). \tag{5.34}$$

in which the intercept term of the logistic regression model, β_{01} , equates the intercept term of the log-linear analysis model, λ_A, indicating the log-odds for A given the condition of $B_i = 0$. Moreover, when $B_i = 1$,

$$\text{logit}(p(A_i = 1 \mid B_i = 1)) = \log\left(\frac{p(A_i = 1 \mid B_i = 1)}{p(A_i = 0 \mid B_i = 1)}\right) = \log\left(\frac{\mathbf{E}[F_{11}]}{\mathbf{E}[F_{01}]}\right)$$

$$= \beta_{01} + \beta_{11} = \lambda_A + \lambda_{AB}. \tag{5.35}$$

$$\Rightarrow \beta_{11} = \log\left(\frac{\mathbf{E}[F_{11}]}{\mathbf{E}[F_{01}]}\right) - \beta_{01} = \log\left(\frac{\mathbf{E}[F_{11}]}{\mathbf{E}[F_{01}]}\right) - \log\left(\frac{\mathbf{E}[F_{10}]}{\mathbf{E}[F_{00}]}\right)$$

$$= \log\left(\frac{\mathbf{E}[F_{11}]\mathbf{E}[F_{00}]}{\mathbf{E}[F_{01}]\mathbf{E}[F_{10}]}\right) = \lambda_{AB}, \tag{5.36}$$

in which the coefficient term of the logistic regression model, β_{11}, equates the interaction term of the log-linear analysis model, λ_{AB}, indicating the log-odds ratio between A and B. Similarly, we evaluate λ_B, β_{02} and β_{12} with a similar method, conditioning on A, and derive the following result:

1. $\lambda_B = \beta_{02}$ denotes the log odds for B given the condition of $A_i = 0$,
2. $\beta_{12} = \beta_{11} = \lambda_{AB}$ denotes the log odds ratio between A and B.

This comparison can be extended to a data set with three variables. Using the same data sets and the pattern table in Section 5.3 with the same variables, A, B and C, the equivalent logistic regression models are defined by the following equations:

$$\text{logit}(p(A_i = 1 \mid B_i, C_i)) = \log\left(\frac{p(A_i = 1 \mid B_i, C_i)}{p(A_i = 0 \mid B_i, C_i)}\right) = \beta_{01} + \beta_{21}B_i + \beta_{31}C_i; \tag{5.37}$$

$$\text{logit}(p(B_i = 1 \mid A_i, C_i)) = \log\left(\frac{p(B_i = 1 \mid A_i, C_i)}{p(B_i = 0 \mid A_i, C_i)}\right) = \beta_{02} + \beta_{12}A_i + \beta_{32}C_i; \tag{5.38}$$

$$\text{logit}(p(C_i = 1 \mid A_i, B_i)) = \log\left(\frac{p(C_i = 1 \mid A_i, B_i)}{p(C_i = 0 \mid A_i, B_i)}\right) = \beta_{03} + \beta_{13}A_i + \beta_{23}B_i. \tag{5.39}$$

The explanation of the relationship between log-linear analysis model and logistic regression model begins by the condition (a): $p(C_i = 1 \mid A_i, B_i)$ for $A_i = 0, B_i = 0$.

In the log-linear analysis model, from Equations 5.10 and 5.12:

$$\log\left(\frac{\mathbf{E}[F_{001}]}{\mathbf{E}[F_{000}]}\right) = \lambda_C. \tag{5.40}$$

Using Equation 5.39 and letting $F_{A_i=a,B_i=b}$ as the frequency of cases fulfilling the two conditions $A_i = a$ and $B_i = b$, the equivalent logistic regression model is expressed as follows:

$$\log\left(\frac{p(C_i=1 \mid A_i=0,B_i=0)}{p(C_i=0 \mid A_i=0,B_i=0)}\right) = \log\left(\frac{\mathbf{E}[F_{001}]/\mathbf{E}[F_{A_i=0,B_i=0}]}{\mathbf{E}[F_{000}]/\mathbf{E}[F_{A_i=0,B_i=0}]}\right) = \beta_{03}. \quad (5.41)$$

As a result, $\lambda_C = \beta_{03}$ represent the log odds for C when $A_i = 0, B_i = 0$.

When evaluating another condition (b): $A_i = 0, B_i = 1$, from Equation 5.23:

$$\lambda_{BC} = \log\left(\frac{\mathbf{E}[F_{011}]\mathbf{E}[F_{000}]}{\mathbf{E}[F_{010}]\mathbf{E}[F_{001}]}\right). \quad (5.42)$$

In the logistic regression model,

$$\log\left(\frac{p(C_i=1 \mid A_i=0,B_i=1)}{p(C_i=0 \mid A_i=0,B_i=1)}\right) = \log\left(\frac{\mathbf{E}[F_{011}]/\mathbf{E}[F_{A_i=0,B_i=1}]}{\mathbf{E}[F_{010}]/\mathbf{E}[F_{A_i=0,B_i=1}]}\right) = \beta_{03}+\beta_{23}. \quad (5.43)$$

$$\Rightarrow \beta_{23} = \log\left(\frac{\mathbf{E}[F_{011}]}{\mathbf{E}[F_{010}]}\right) - \beta_{03} = \log\left(\frac{\mathbf{E}[F_{011}]}{\mathbf{E}[F_{010}]}\right) - \log\left(\frac{\mathbf{E}[F_{001}]}{\mathbf{E}[F_{000}]}\right)$$

$$= \log\left(\frac{\mathbf{E}[F_{011}]\mathbf{E}[F_{000}]}{\mathbf{E}[F_{001}]\mathbf{E}[F_{010}]}\right) = \lambda_{BC}. \quad (5.44)$$

According to the result of Equation 5.44, $\lambda_{BC} = \beta_{23}$ represents the log-odds ratio between B and C when $A_i = 0$.

Finally, when evaluating another condition (c): $A_i = 1, B_i = 0$, from Equation 5.22:

$$\lambda_{AC} = \log\left(\frac{\mathbf{E}[F_{101}]\mathbf{E}[F_{000}]}{\mathbf{E}[F_{100}]\mathbf{E}[F_{001}]}\right). \quad (5.45)$$

In the logistic regression model, using Equations 5.39,

$$\log\left(\frac{p(C_i=1 \mid A_i=1,B_i=0)}{p(C_i=0 \mid A_i=1,B_i=0)}\right) = \log\left(\frac{\mathbf{E}[F_{101}]/\mathbf{E}[F_{A_i=1,B_i=0}]}{\mathbf{E}[F_{100}]/\mathbf{E}[F_{A_i=1,B_i=0}]}\right) = \beta_{03}+\beta_{13}. \quad (5.46)$$

$$\Rightarrow \beta_{13} = \log\left(\frac{\mathbf{E}[F_{101}]}{\mathbf{E}[F_{100}]}\right) - \beta_{03} = \log\left(\frac{\mathbf{E}[F_{101}]}{\mathbf{E}[F_{100}]}\right) - \log\left(\frac{\mathbf{E}[F_{001}]}{\mathbf{E}[F_{000}]}\right).$$

$$= \log\left(\frac{\mathbf{E}[F_{101}]\mathbf{E}[F_{000}]}{\mathbf{E}[F_{001}]\mathbf{E}[F_{100}]}\right) = \lambda_{AC}. \tag{5.47}$$

According to the result of Equation 5.47, $\lambda_{AC} = \beta_{13}$ represents the log-odds ratio between A and C when $B_i = 0$.

On the other hand, conditioning on $A_i = 1$, the log-odds ratio between B and C is expressed as follows:

In the log-linear analysis model, (5.17) + (5.14) - (5.15) - (5.16),

$$log\left(\frac{\mathbf{E}[F_{111}]\mathbf{E}[F_{100}]}{\mathbf{E}[F_{101}]\mathbf{E}[F_{110}]}\right) = (\lambda_0 + \lambda_A + \lambda_B + \lambda_C + \lambda_{AB} + \lambda_{AC} + \lambda_{BC}) + (\lambda_0 + \lambda_A)$$

$$-(\lambda_0 + \lambda_A + \lambda_C + \lambda_{AC}) - (\lambda_0 + \lambda_A + \lambda_B + \lambda_{AB}) = \lambda_{BC}. \tag{5.48}$$

In the logistic regression model, using Equation 5.39,

$$\log\left(\frac{p(C_i = 1 \mid A_i = 1, B_i = 1)}{p(C_i = 0 \mid A_i = 1, B_i = 1)}\right) = \beta_{03} + \beta_{13} + \beta_{23}, \tag{5.49}$$

$$\log\left(\frac{p(C_i = 1 \mid A_i = 1, B_i = 0)}{p(C_i = 0 \mid A_i = 1, B_i = 0)}\right) = \beta_{03} + \beta_{13}. \tag{5.50}$$

(5.49) - (5.50),

$$\log\left(\frac{p(C_i = 1 \mid A_i = 1, B_i = 1)}{p(C_i = 0 \mid A_i = 1, B_i = 1)}\right) - \log\left(\frac{p(C_i = 1 \mid A_i = 1, B_i = 0)}{p(C_i = 0 \mid A_i = 1, B_i = 0)}\right)$$

$$= \log\left(\frac{(\mathbf{E}[F_{111}]/\mathbf{E}[F_{A_i=1,B_i=1}])}{(\mathbf{E}[F_{110}]/\mathbf{E}[F_{A_i=1,B_i=1}])}\right) - \log\left(\frac{(\mathbf{E}[F_{101}]/\mathbf{E}[F_{A_i=1,B_i=0}])}{(\mathbf{E}[F_{100}]/\mathbf{E}[F_{A_i=1,B_i=0}])}\right)$$

$$= \log\left(\frac{\mathbf{E}[F_{111}]\mathbf{E}[F_{100}]}{\mathbf{E}[F_{101}]\mathbf{E}[F_{110}]}\right) = \beta_{23} = \lambda_{BC}, \tag{5.51}$$

implying that $\lambda_{BC} = \beta_{23}$ represents the log-odds ratio between B and C when $A_i = 0, 1$, in other words, in all conditions of variable A.

Similarly, (5.17) + (5.11) - (5.13) - (5.15)

$$\log\left(\frac{\mathbf{E}[F_{111}]\mathbf{E}[F_{010}]}{\mathbf{E}[F_{011}]\mathbf{E}[F_{110}]}\right) = (\lambda_0+\lambda_A+\lambda_B+\lambda_C+\lambda_{AB}+\lambda_{AC}+\lambda_{BC})+(\lambda_0+\lambda_B)$$
$$-(\lambda_0+\lambda_B+\lambda_C+\lambda_{BC})-(\lambda_0+\lambda_A+\lambda_B+\lambda_{AB})=\lambda_{AC}. \quad (5.52)$$

In the logistic regression model, using Equation 5.39,

$$\log\left(\frac{p(C_i=1\mid A_i=1,B_i=1)}{p(C_i=0\mid A_i=1,B_i=1)}\right)=\beta_{03}+\beta_{13}+\beta_{23}, \quad (5.53)$$

$$\log\left(\frac{p(C_i=1\mid A_i=0,B_i=1)}{p(C_i=0\mid A_i=0,B_i=1)}\right)=\beta_{03}+\beta_{23}. \quad (5.54)$$

(5.53) - (5.54),

$$\log\left(\frac{p(C_i=1\mid A_i=1,B_i=1)}{p(C_i=0\mid A_i=1,B_i=1)}\right)-\log\left(\frac{p(C_i=1\mid A_i=0,B_i=1)}{p(C_i=0\mid A_i=0,B_i=1)}\right)$$
$$=\log\left(\frac{\mathbf{E}[F_{111}]\mathbf{E}[F_{010}]}{\mathbf{E}[F_{011}]\mathbf{E}[F_{110}]}\right)=\beta_{13}=\lambda_{AC}, \quad (5.55)$$

implying that $\lambda_{AC}=\beta_{13}$ represents the log odds ratio between variables A and C when $B_i=0,1$, in other words, in all conditions of variable B.

Applying similar derivation techniques on Equations 5.10 to 5.17, and using Equations 5.37 and 5.38, the following results are generated:

$\lambda_A=\beta_{01}$ represents the log-odds for A when $B_i=0,C_i=0$;

$\lambda_B=\beta_{02}$ represents the log-odds for B when $A_i=0,C_i=0$;

$\lambda_C=\beta_{03}$ represents the log-odds for C when $A_i=0,B_i=0$;

$\lambda_{AC}=\beta_{13}=\beta_{31}$ represents the log-odds ratio between A and C;

$\lambda_{AB}=\beta_{12}=\beta_{21}$ represents the log-odds ratio between A and B;

$\lambda_{BC}=\beta_{23}=\beta_{32}$ represents the log odds ratio between B and C.

In general, intercept terms of the log-linear analysis models represent the log-odds of the respective drug-trying response variables, under the condition that all other responses be "0". Also, interaction terms of the log-linear analysis models represent the log-odds ratios between two drug-trying response variables in the data set.

To compare the log-linear analysis models with the univariate logistic regression models for drug-trying response variables only, the same corresponding combinations of covariate terms to all interaction terms in the saturated log-linear analysis model are adopted in the saturated univariate logistic regression models. In other words, whenever an interaction term is not included in a log-linear analysis model, the corresponding combinations of response variables and covariates are not included in the univariate logistic regression models.

In this section, the comparison of the interaction terms of the saturated log-linear analysis model with the terms of the saturated univariate logistic regression models is made. The log-odds ratio heat plots for the saturated models are presented in Figures 5.2 and 5.7 respectively and the related covariate sign plots were are presented in Figures 5.4 and 5.8 respectively.

When comparing Figures 5.2 and 5.7, both saturated models of the log-linear analysis model and the univariate logistic regression models (after backward elimination) exhibited the dominance of cannabis in terms of relationships with other types of drugs, showing that the students trying cannabis were more likely to try other types of drugs, or vice versa. From Figures 5.4 and 5.8, a majority of the significant interaction terms in both models showed positive associations among drug-trying response variables, however, less negative interaction terms were found in the univariate logistic regression models than in the log-linear

analysis models. Most interaction terms in both models were positive interaction terms, such as those involving ecstasy and cannabis as well as heroin and cocaine, of which indicated positive associations between such pair of drugs.

5.4 Summary

In this chapter, both univariate logistic regression models and log-linear analysis models were applied to further explore possible interactions among drug-trying response variables and to understand the associations of the smoking, drinking and drug-related socio-demographic covariates with students' drug-trying behaviour.

The univariate logistic regression models reported the one-way interaction among the fifteen drug-trying response variables. The main findings reflected by the univariate logistic regression models included:

1. Almost every drug has positive interactions with other types of drugs.

2. The extent of interactions among drug-trying response variables varied among the fifteen drugs.

3. Among the fifteen drugs, cannabis was found positively associating with the highest number of other types of drugs. On the other hand, methadone, crack and other drugs were found associating with a relatively smaller number of other types of drugs.

The log-linear analysis models reported the two-way interaction among the fifteen drug-trying response variables. Apart from that, the results from the

saturated log-linear analysis models were found generally comparable with that of the univariate logistic regression models, particularly in the following two dimensions:

1. A large number of significant interaction terms, in terms of log-odds ratios, between all drugs were found, and most of these interaction terms were positive, with only a few being negative.
2. Among the fifteen drugs, cannabis was the most dominant drug that yielded the greatest number of significant interaction terms with other types of drugs.

The univariate logistic regression models further revealed that numerous smoking, drinking and drug-related socio-demographic covariates were associated with students' drug-trying behaviour in different extent. These covariates replaced several drug covariates in predicting whether a student had ever tried at least one of the fifteen drugs. These covariates were summarised as follows:

Smoking covariates included: (1) family attitudes toward smoking (CgFam1); (2) cigarette smoking status (CgStat); (3) sources of purchasing cigarettes (CgGet); (4) number of smokers who the students know and where those smokers smoked (CgWho1) and (5) education and information about smoking (CgPe1 and CgIn1).

Drinking covariates included: (1) time and frequency of consuming alcohol (AlFreq2); (2) number of alcohol drinkers students know and where those drinkers drank (AlEstim, AlBnPub); (3) family's attitude towards drinking (AlPar1); (4) how students purchase alcohol (AlBuy1, AlBuy2, AlBuy) and where they consume the alcohol (AlUs1, AlUs2); (5) having lessons or obtaining information about drinking (AlPe1, AlIn1) and (6) types of issues happened when a student drank alcohol (Al4W1).

Drug-related socio-demographic covariates included: (1) having lessons or obtaining information about drugs (DgPe1, DgIn1); (2) number of drug-trying students know and where those drug-takers tried drugs (DgEstim) (3) the amount of books students possessed (Books1); (4) age; (5) gender; (6) free school meal scheme; (FSM1) (7) frequency of truancy (TruantN); (8) frequency of being excluded (ExclAN1) and (9) SHA (SHA).

Both the univariate logistic regression models and the log-linear analysis models have shown a large number of covariates predicting students' drug-trying behaviour. This finding is useful for latent class regression modelling. As a large number of interaction terms between drug-trying response variables were detected, this finding supports the feasibility of analysing multiple drug-trying response variables in a single item response theory model, a single latent class analysis model and through k-means clustering. We will discuss the fitting of the item response theory model on our working data set in Chapter 6 as well as the running of the latent class analysis and k-means clustering in Chapter 7.

Chapter 6

Item Response Theory

In this chapter, we discuss the use of the item response theory model for the investigation of drug use among young people, as well as the key feature of the item response theory model and the Item Characteristic Curve (ICC) (Hambleton et al., 1991), which is a logistic curve for the probability of a positive response under different values of a latent parameter (Loken and Rulison, 2010). In particular, we conduct this investigation in order to understand more about how each drug-trying response variable relates to the overall drug-trying behaviour of the students, as well as the proportion of students trying each drug. We commence with a brief overview of the underpinning item response theory prior to fitting an item response theory model to the working data set. We also compare the likelihood and Bayesian approaches and contrast them in terms of their statistical inferences.

6.1 Introduction

Both Lord (1951) and Rasch (1960) have laid down a solid foundation on early work of the item response theory model. Lord (1951) adopted a large number of item response theory terms such as "latent ability", which means there is a hidden parameter that explains the "ability" parameter of the respondents in

the data set, whereas Rasch (1960) published the one-parameter item response theory model. Birnbaum (1967) developed the model further, forming the two-parameter and three-parameter item response theory models. Based upon early works from Lord (1951), Rasch (1960) and Birnbaum (1967), the item response theory was utilised by Lord (1968) to model the data of the Verbal Scholastic Aptitude Test (Verbal SAT). The consequent statistical model, the item response theory model, was used to analyse students' performance in different SAT tests, in terms of separation scores among students, namely a scale parameter. Also, it was used to analyse the amount of influence of trying a certain drug on overall drug-trying behaviour, with a "discrimination" parameter, as well as measuring the drug's location on the scale that quantified the proportion of respondents trying the drug, with a "difficulty" parameter. Over the last three decades, the item response theory model has been universally adopted in "psychometrics and educational measurements" (Carlson and von Davier, 2013), and it is in continuous evolution. The model has also been increasingly popular in the social science and education sectors, and its application has been extended to other domains, such as investigation of personality (Reise and Waller (1990); Ferrando (1994); Rouse et al. (1999)) and delinquency (Osgood et al., 2002).

In this research, the item response theory model is considered appropriate in the investigation of the 15 drug-trying response variables as it will allow for analysing the probability that a student tries a drug, as well as the separation amongst students regarding their drug-trying behaviour. The item response theory model can also explain the proportion of the students trying each drug, whilst providing additional information about the degree of separation of each drug-trying response variable and the drug-trying behaviour of the students.

In this research, we adopted the two-parameter item response theory model to further investigate the relationships between drug-trying response variables

and the students' overall drug-trying behaviour. As described in Arima (2015), the item response theory model differ from the conventional univariate logistic regression model and the log-linear analysis model that the item response theory model is based on the "invariance property": (1) parameters that characterise the drug items do not depend on how the likelihoods of students to try each drug are distributed and (2) parameters that characterise a student do not depend on the drug responses. In addition, the item response theory models are based on the following three postulates: (1) the likelihood of students to try a drug can be explained by a latent parameter (unidimensionality); (2) the observed drug responses are conditionally independent of each other, given the latent parameter that measures the overall likelihood of each respondent to try drugs, and (3) the relationship between the likelihood of students to try a drug and the overall likelihood of each student to try drugs can be described by the item characteristic curve (ICC). Based on the above characteristics of the item response theory model, it can help to investigate the influence of trying each drug by the students on their entire drug-trying behaviour, which is not found in the log-linear analysis and univariate logistic regression models, other than measuring the likelihood for students to try certain drugs. Such finding can support the results of log-odds of trying each drug in the log-linear analysis and univariate logistic regression models.

This introduction provided a brief description of the item response theory model. In the next section, we discuss the theory of the item response theory model in more details.

6.2 Theory

The item response theory model is used for evaluating the probability of respondent $i, i = 1, \ldots, n$, to make a positive binary response $(Y_{ij} = 1)$ when presented

with item $j, j = 1, \ldots, J$. Such a probability is denoted as $P(Y_{ij} = 1)$. The factor level or ability parameter for the respondent i, denoted as θ_i, measures the latent tendency of the respondent i to yield a positive response.

By letting ζ_j as a collection of parameters that describe the characteristics of item j, the general form of the item response theory model (for binary responses) is expressed as:

$$P(y_{ij} = 1 \mid \theta_i, \zeta_j)) = f(y_{ij} \mid \theta_i, \zeta_j) \tag{6.1}$$

(Baker, 1961).

The Rasch model (also known as the one-parameter Rasch model) only contains the factor score parameter (θ_i) for respondent i, as well as a difficulty factor (δ_j) for item j, hence in this case, $\zeta_j = \{\delta_j\}$. The equation for the Rasch model is expressed as:

$$\text{logit}(P(y_{ij} = 1 \mid \theta_i, \delta_j)) = (\theta_i - \delta_j) \tag{6.2}$$

(Rasch, 1960).

The two-parameter item response theory model contains the factor score parameter (θ_i) for respondent i, as well as a discrimination factor (α_j) and a difficulty factor (δ_j) for item j, hence in this case, $\zeta_j = \{\alpha_j, \delta_j\}$. A discrimination parameter for item j, denoted as α_j, measures how well the item j separates the respondents. The difficulty parameter of the item j, denoted as δ_j, measures the difficulty of the item j. α_j and δ_j are referred to as fixed effects, whereas θ_i is referred to as a random effect. In general, the two-parameter item response theory model is expressed as:

$$\text{logit}(P(y_{ij} = 1 \mid \theta_i, \alpha_j, \delta_j)) = \alpha_j(\theta_i - \delta_j) \tag{6.3}$$

(Rizopoulos, 2006).

Equation 6.3 can be expressed alternatively as:

$$P(y_{ij} = 1 \mid \theta_i, \alpha_j, \delta_j) = \frac{\exp(\alpha_j(\theta_i - \delta_j))}{1 + \exp(\alpha_j(\theta_i - \delta_j))}. \tag{6.4}$$

(Van der Linden et al., 1997)

The comparison between the Rasch model and the two-parameter item response theory model is illustrated in Figure 6.1.

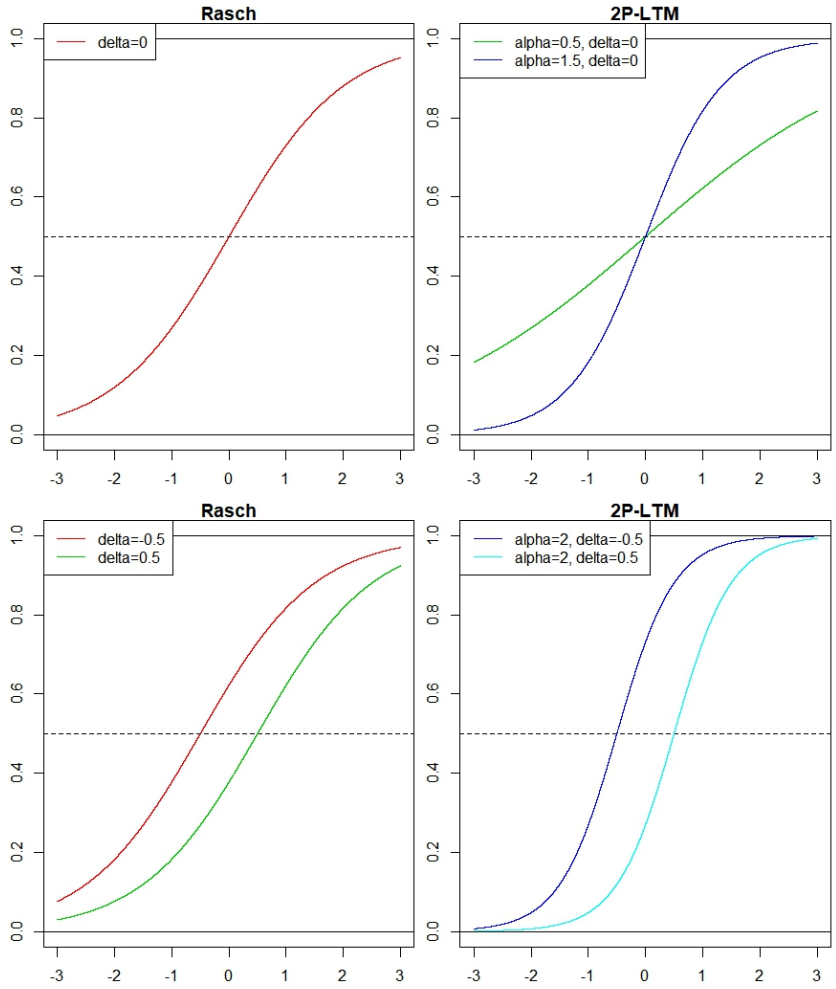

Figure 6.1: Comparison of Item Characteristic Curves between Rasch Model and the Two-parameter Item Response Theory model with varied Discrimination factor (fixed to 1 for Rasch model) (Upper pair) and Difficulty factor (factor of value -0.5 versus 0.5) (Lower pair) (x-axis: difficulty factor value, y-axis: factor score value).

As observed from Figure 6.1, when the discrimination factor was varied for two-

parameter item response theory model, the item characteristic curve was flatter as the discrimination factor was smaller and was steeper as the discrimination factor was larger. When the difficulty factor was varied, the item characteristic curve shifted leftwards as the difficulty factor decreased, and rightwards as the difficulty factor increased.

The reason we employed two-parameter item response theory model rather than one-parameter item response model was because from the results of univariate logistic regression and log-linear analysis models, different drug-trying response variables possessed different relationships between each other drug-trying response variable, and various frequencies of drug-trying patterns were measured, implying that different drug-trying variables may influence the entire drug-trying behaviour of the students at different degree. As such, a varying discrimination parameter was required. We also employed two-parameter item response theory model rather than three-parameter item response model based on the following reasons: (1) the third parameter of the three-parameter item response model is the guessing parameter, which may potentially predict students' probability of trying the right drug, which may not be reasonable in this analysis, and (2) the correct model for three-parameter item response model remains uncertain (von Davier, 2009).

When fitting an item response theory model, there are two possible approaches, namely the marginal approach and the Bayesian approach, that can be adopted and are implemented in `R` program. These two approaches will be discussed in Sections 6.2.1 and 6.2.2 respectively.

6.2.1 Rizopoulus Marginal Approach

In the marginal approach, both the two-parameter item response theory model and the Rasch model are fitted by marginal maximum likelihood approach. In the marginal maximum likelihood estimation, model parameters, namely discrimination factor and difficulty factor (α_j, δ_j) are estimated through integrating the latent variables or random effect (θ_i) out of the equation, obtaining marginal log-likelihood and then maximising this marginal log-likelihood (Rizopoulos, 2006).

Let $\mathbf{y}_i$ be the vector of responses for the i^{th} individual respondent, and α_j and δ_j be the discrimination and difficulty factors for item j of an item response theory model respectively. Assume that the latent parameter for the i^{th} respondent, θ_i, follows a standard normal distribution, then the log-likelihood equation for i^{th} respondent is:

$$l_i(\alpha_j, \delta_j) = \log\left(p(y_i; \alpha_j, \delta_j)\right) = \log \int p(y_i \mid \theta_i; \alpha_j, \delta_j) p(\theta_i) d\theta_i \tag{6.5}$$

(McCullagh and Nelder, 1999).

Considering the parameters of an item response theory model, the likelihood for i^{th} respondent, which is conditioned on the latent parameter-factor score is expressed as follows:

$$\mathrm{L}_i(\alpha_j, \delta_j; \theta_i) = p(y_i \mid \theta_i; \alpha_j, \delta_j) = \prod_j \left\{ \frac{\exp(\alpha_j(\theta_i - \delta_j))}{1 + \exp(\alpha_j(\theta_i - \delta_j))} \right\} \tag{6.6}$$

$$= \frac{\exp(\sum_j \alpha_j \theta_i - \sum_j \alpha_j \delta_j)}{\prod_j [1 + \exp(\alpha_j(\theta_i - \delta_j))]} \tag{6.7}$$

(Rizopoulos, 2006).

Rizopoulos (2006) made an assumption that $\theta_i \sim \text{Normal}(0,1)$, then for $i = 1,\ldots,n$, the marginal log-likelihood is as follows:

$$l_i(\alpha_j,\delta_j) = \log \int_i \frac{\exp(\sum_j \alpha_j\theta_i - \sum_j \alpha_j\delta_j)}{\prod_j[1+\exp(\alpha_j(\theta_i-\delta_j))]} \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\theta_i^2)d\theta_i. \tag{6.8}$$

In Equation 6.8, the factor score component cannot be integrated out analytically but can be approximated by Gauss-Hermite quadrature.

The likelihood optimisation through Hessian matrix is initially processed by the expectation-maximisation (EM) updating algorithm until convergence (Rizopoulos, 2006).

The EM updating algorithm (Dempster et al., 1977) intends to evaluate parameters α_j and δ_i, which maximises the likelihood optimization by two steps: an expectation step (E-step), which computes the expected likelihood function for these three parameters. The E-step is followed by a maximisation step (M-step), which finds the respective values of α_j and δ_j that maximises the likelihood function. The EM steps are conducted for T iterations. For iteration t for $t = 1,\ldots,T$, E-step and M-step are listed as follows:

E-step: Compute $\mathbf{E}[l_i(\alpha_j^{(t)},\delta_j^{(t)})]$;
M-step: Evaluate $\left(\alpha_j^{(t+1)},\delta_j^{(t+1)}\right) = \text{argmax}\left(\mathbf{E}\left[l_i\left(\alpha_j^{(t)},\delta_j^{(t)}\right)\right]\right)$
then set $t = t+1$

(Dempster et al., 1977).

In this research, estimates and standard deviations of parameters, for both discrimination factors and difficulty factors, were combined and evaluated by Rubin's rule, which was described in Section 4.5.3.3, to provide corresponding pooled estimates and standard errors across all imputed data sets, accounting

for imputation uncertainty. To evaluate the factor scores θ of the item response theory model, the empirical Bayes estimation method was adopted.

An alternative to the marginal approach of Rizopoulos (2006) is the Bayesian approach, when the factor score, θ, is not integrated out, but instead being estimated along with discrimination and difficulty factors.

6.2.2 Bayesian Approach with OpenBUGS

The Bayesian approach adopts the Markov Chain Monte Carlo (MCMC) algorithm (Metropolis et al. (1953); Hastings (1970)). In the Bayesian approach, priors for discrimination factor, difficulty factor and factor scores are specified. Posterior mean estimates, as well as standard deviations of discrimination factors, difficulty factors and factor scores are generated after updating of values through Markov chains. In Bayesian inference, the priors of a set of parameters θ, α, δ, denoted as $p(\theta)$, $p(\alpha)$, $p(\delta)$ respectively, must be specified to generate chains of iterations for posterior inference. The joint probability density function of parameters α, δ and θ, with conditioning on data response y is denoted as $p(\alpha, \delta, \theta \mid y)$. The posterior is the probability density function of a parameter conditioned on data response, i.e. $p(\alpha, \delta, \theta \mid y)$. The posterior is evaluated using the Bayes Theorem. It is proportional to likelihood and prior, which is expressed in the following equation.

$$\begin{aligned} p(\alpha, \delta, \theta \mid y) &\propto p(y \mid \alpha, \delta, \theta) p(\alpha, \delta, \theta) \\ &\propto p(y \mid \alpha, \delta, \theta) p(\alpha) p(\delta) p(\theta). \end{aligned}$$

Algorithm 6.1 below shows the Gibbs Sampling procedure for finding the posterior distributions of θ, α and δ, in the presence of missing data. Algorithm 6.1 of Bayesian approach for an item response theory model is described as:
The entire procedure in lines 1 to 15 of algorithm 6.1 is repeated until convergence

Algorithm 6.1 Bayesian Approach

for Chains $c = 1, \ldots C$ **do**
  Intialise all unknown parameters from the full posterior conditionals
  $\{\mathbf{y}_m, \theta, \delta, \alpha\} \leftarrow \{y_m^{(0)}, \theta^{(0)}, \delta^{(0)}, \alpha^{(0)}\}$
  for $t = 1, \ldots, T$ **do**
    Missing data is simulated from likelihood: $y_{miss} \mid \ldots \sim p(\mathbf{y} \mid \theta, \alpha, \delta)$.
    for $j = 1, \ldots, P$ **do**
      The unknown parameters α_j, δ_j are sampled for chains $1 : C$;
      $\alpha_j \mid \ldots \sim p(\alpha_j)p(y \mid \theta, \alpha, \delta)$ - to sample from this distribution, propose $q(\alpha_j)$ and accept according to Metropolis-Hastings ratio;
      $\delta_j \mid \ldots \sim p(\delta_j)p(y \mid \theta, \alpha, \delta)$ - to sample from this distribution, propose $q(\delta_j)$ and accept according to Metropolis-Hastings ratio.
    end for
    for $i = 1, \ldots, n$ **do**
      $\theta_i \mid \ldots \sim p(\theta_i)p(y \mid \theta, \alpha, \delta)$.
    end for
  end for
end for

of all parameters (i.e. θ, α, δ) in the model. In algorithm 6.1, the notation $\alpha_j \mid \ldots$ denotes the full conditional posterior distribution of α_j given everything else except α_j and the burn-in iterations (see Section 6.2.3 below). The same notation applies for δ_j (i.e. $\delta_j \mid \ldots$).

6.2.3 Comparison of Bayesian Approach to Marginal Approach

The advantages of a Bayesian approach over a marginal approach are: (1) missing data are updated along with other parameters in the one-stage model instead of two-stage model, where the imputation model and substantive model are separated instead to being integrated into a single model; (2) the process of generating iterations is faster and (3) creating multiple data sets is not required. On the other hand, Bayesian approach has several drawbacks: (1) under the Bayesian approach, it is much more difficult to select an appropriate imputation model, when there are covariates with missing values; (2) sensible priors should be used for all parameters, and their sensitivities should be tested and (3) large computational power is required to model a large number of variables in

Bayesian approach, and thus making it less feasible for analysis involving many variables.

6.3 Application of Item Response Theory Models

In order to investigate how the drug-trying response variables discriminate and differ in terms of their discrimination and difficulty factors, the item response theory model was implemented on the working data set through two approaches: (1) the marginal approach and (2) the Bayesian approach.

The marginal approach (using `ltm` package in `R` program) (Rizopoulos, 2006) was explored by using two different schemes for handling missing data: (1) MICE scheme on 15 drug-trying response variables only (Scheme 1) and (2) MICE scheme on full data set (Scheme 2). In the Bayesian approach, the item response theory model was applied to data sets imputed through the `OpenBUGS` program.

6.3.1 Marginal Approach

In the marginal approach, the 15 drug-trying response variables imputed from each of the two schemes of imputed data sets were analysed with complete case analysis. In the complete case analysis, 6,791 students were involved in the model fitting and statistical inference. For the two MICE schemes and the complete case analysis, the `ltm` package in `R` program was adopted for model fitting, with 21 points of Gauss-Hermite estimation.

6.3.2 Bayesian Approach

In the Bayesian approach, a range of discrimination and difficulty prior specifications were adopted in the item response theory model through specifying

a specific item response theory model in `OpenBUGS` program, spanning 14 prior specifications for the discrimination prior and two prior specifications for the difficulty prior.

Bazan et al. (2006) used the half normal priors for the discrimination factor. Both Patz and Junker (1999) and Sahu (2002) suggested log-normal distribution for the discrimination factor. Glickman et al. (2009) suggested norm (0, 100) distribution for the difficulty factor. Norm (0, 1) (Glickman et al., 2009) was adopted as the prior for factor score (θ) The following priors for discrimination factor (α), difficulty factor (δ) were adopted in Table 6.3.1. The distribution plot for priors of discrimination factor is presented in Figure 6.2.

Table 6.3.1: Table of Priors for Parameters in OpenBUGS

Prior	α
1	Gamma (1, 0.1)
2	Gamma (1, 0.311) (Roos and Held, 2011)
3	Gamma (1, 0.622) (Roos and Held, 2011)
4	Gamma (1, 0.933) (Roos and Held, 2011)
5	Half-normal (0, 100) (Ames, 2015)
6	Half-normal (0, 1000)
7	Half-normal (0, 0.5)
8	Log-normal (0, 0.16)
9	Log-normal (1, 0.25)
10	Log-normal (1, 0.5)
11	Uniform (0, 100)
12	Log-normal (0, 4) (Hsieh and Proctor, 2010)
13	Log-normal (0, 0.0625) (Nering and Ostini, 2010)
14	Log-normal (1, 4)
Prior	δ
1	Norm (0, 100) (Glickman et al., 2009)
2	Norm (0, 1000)

Some priors from Figure 6.2 were flat and non-informative and some others were moderately flat but non-informative, such as log-normal (0, 4) and gamma (0, 0.933). Only two of them, log-linear (0, 0.16) and log-linear (1,4), were informative priors. The inclusion of both non-informative and informative priors in the sensitivity analysis for Bayesian approach was to investigate whether the item response theory model result was largely affected by choice of priors, whether

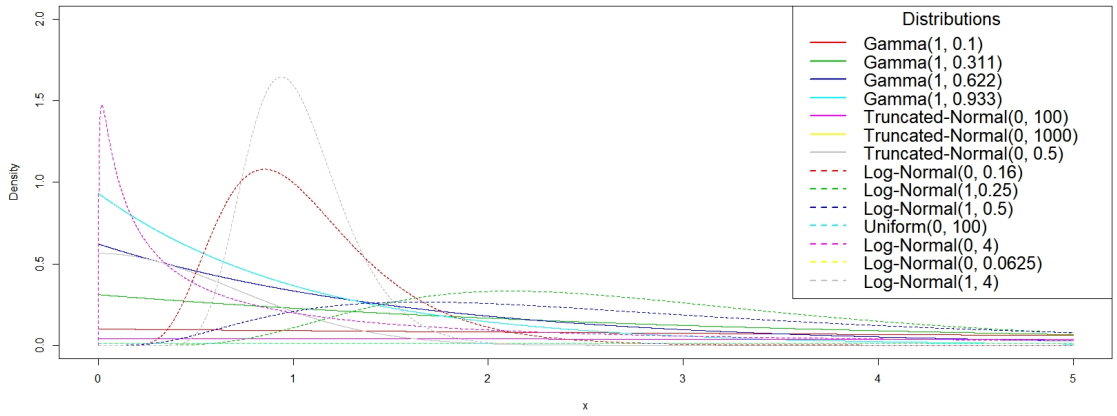

Figure 6.2: Density Plot for Distributions of 14 Priors of Discrimination Factor. For details about prior distributions, please refer to Table 6.3.1.

informative or not.

When fitting an item response theory model in `OpenBUGS` program, the two simulated chains, each contained 17,000 iterations including 1000 burn-in iterations, were generated for each item response theory model. For the first chain, the initial values for all α parameters and δ parameters were set at 1, whilst those for all θ parameters were set at -0.1. For the second chain, the initial values for all α parameters were set at 0.1, whereas those for the δ parameters were set at 0 and those for all θ parameters were set at 0.1. The initial values for missing values of the drug-trying response variables were generated by `OpenBUGS` program with seed number 1 of 14, thereby generating two different sets of initial values. We used two different sets of priors for two respective chains with different initial values for every parameter, in order to diagnose whether the convergence of all parameters has been achieved for a combination of priors. All simulations were processed using seed number 1. The first chain was used for statistical inference, which utilised a total of 16,000 iterations (without the 1,000 burn-in iterations).

The convergence of parameters was diagnosed through trace plots. Examples of these are shown as in Figure 6.6 in Section 6.5. In OpenBUGS program, two chains were executed. For all parameters, if both chains intermingled for a long streak of iterations at a stationary mean, then the convergence was reasonable (Spiegelhalter, 2003). The section before the convergence was called burn-in section when two chains were either not stable or obviously separated from each other. All the iterations during the burn-in section were discarded before any statistical inference.

In Sections 6.4 and 6.5, the results (estimates and standard errors) from the imputed data sets for item response theory models, through two MICE schemes, are firstly discussed by referring to Tables 6.4.1 and 6.4.2. Complete case analysis was adopted in the item response theory analysis, and the subsequent results were included along with those for two MICE schemes as a reference. Secondly, the results of item response theory models from the `OpenBUGS` program were discussed. Then, the item response theory model results generated by `ltm` function in `R` program and the pre-defined item response theory model in `OpenBUGS` program were compared.

Both Sections 6.4 and 6.5 commence with discussion of the 95% confidence interval and mean estimate plots for discrimination and difficulty factors. Finally, we discuss the item characteristic curve, which illustrated the relationship between factor level and the probability of a student trying a specific drug.

6.4 Results of Item Response Theory Model under Marginal Approach

The plots of the pooled estimates and their corresponding 95% confidence intervals from two approaches to imputation and complete case analysis of data sets, followed by tables of combined estimates and standard errors and item characteristic curve plots, are presented in Figures 6.3 and 6.4, as well as in Tables 6.4.1 and 6.4.2 respectively. Ranks from the smallest estimate to the largest estimate were included in Tables 6.4.1 and 6.4.2 and were compared within two approaches to imputation and complete case analysis for their similarities in ordering.

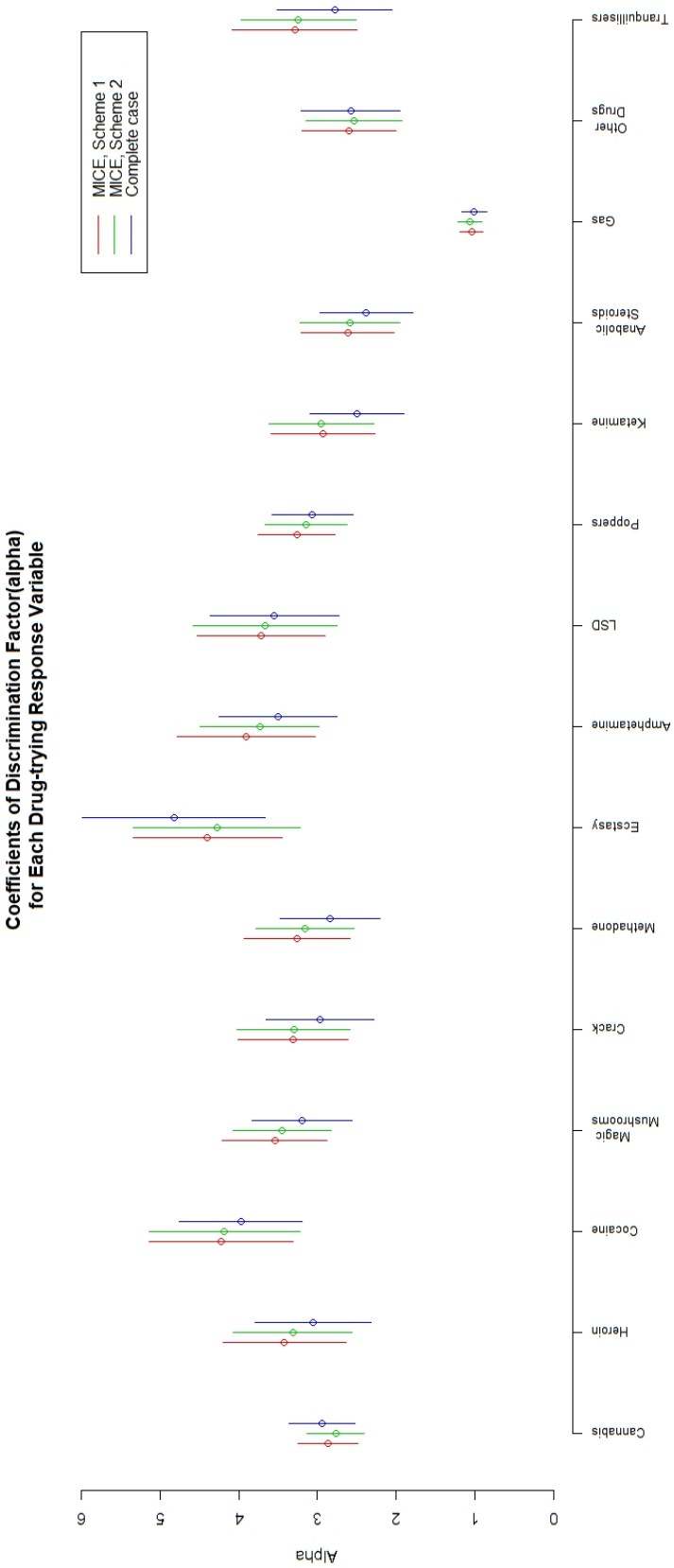

Figure 6.3: Estimate Plots for Two Imputation Schemes and Complete case Analysis of Item Response Theory Model for Discrimination Factor. Red represents MICE Scheme 1, green represents MICE Scheme 2 and blue represents complete case analysis. We noted that the complete case analysis gives a small bias in most cases, both upwards and downwards.

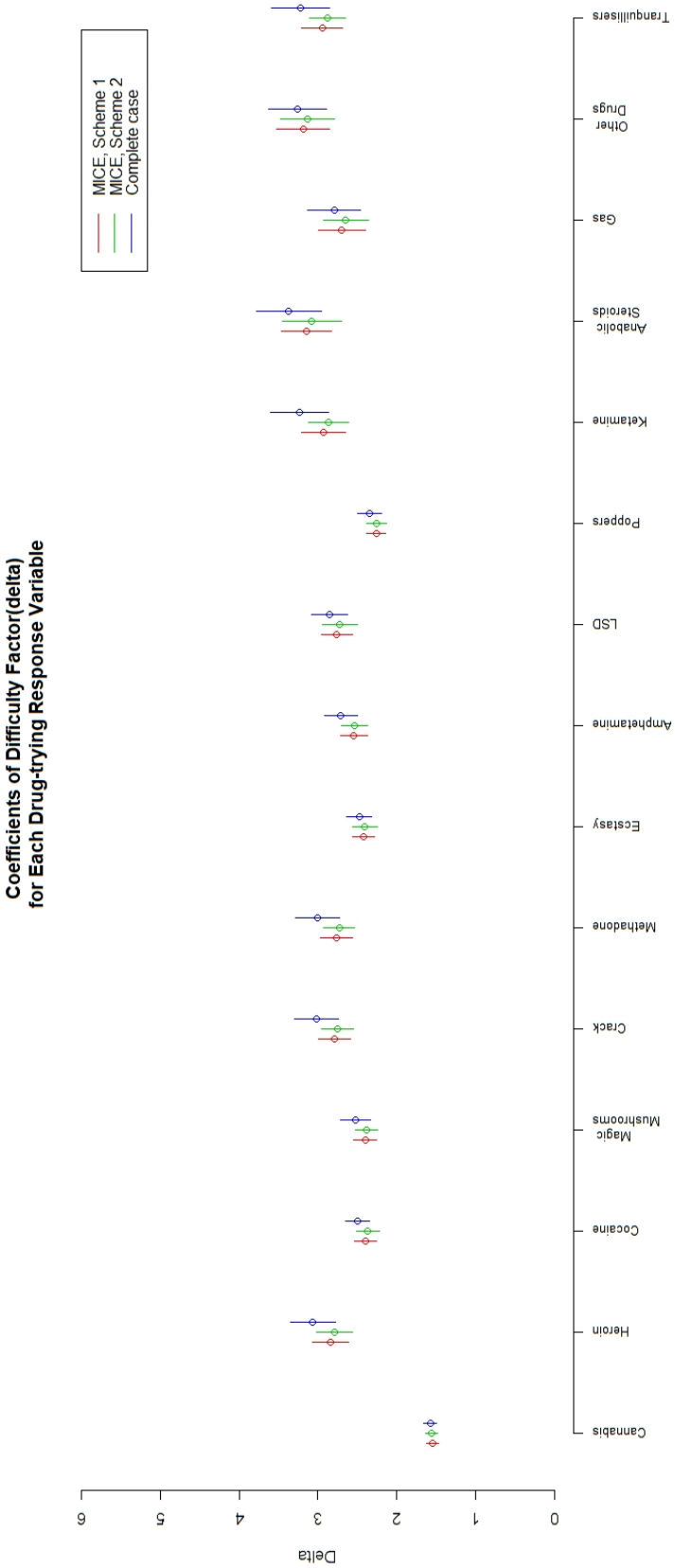

Figure 6.4: Estimate Plots for Two Imputation Schemes and Complete Case Analysis of Item Response Theory Model for Difficulty Factor. Red represents MICE Scheme 1, green represents MICE Scheme 2 and blue represents complete case analysis. We noted that the complete case analysis gives a small bias in most cases, both upwards and downwards.

From Figures 6.3 and 6.4, there were slight differences between the estimates calculated in the complete case analysis and the pooled estimates computed using the imputed data sets and Rubin's rule. The estimates of the discrimination factors, α, calculated in the complete case analysis, appeared to be slightly shifted downwards from the pooled estimates computed using imputed data sets and Rubin's rule. The estimates of the difficulty factors, δ, calculated in the complete case analysis, were slightly shifted upwards from the pooled estimates computed using imputed data sets and Rubin's rule. These slight differences might be caused by the changes in the mean values of drug-trying response variables, due to positive correlation within drug-trying response variables, as well as between drug-trying response variables and most smoking, drinking and socio-demographic covariates. The MICE imputation trace plots in Figures 4.9 to 4.11 in Section 4.7.3 previously showed how the mean values of each drug-trying response variables were influenced by the other variables during 200 iterations of MICE imputation. Though most mean values of the drug-trying response variables were close to the initial values, several of them were different to their initial values.

For discrimination factor, gas yielded the smallest estimate. On the contrary, ecstasy yielded the largest estimate. For difficulty factor, cannabis yielded the smallest estimate, whereas other drugs yielded the largest estimate. This corresponds to the fact that the highest proportion of the students had tried cannabis.

Discrimination Factor						
	Complete case		MICE Scheme 1		MICE, Scheme 2	
Drug	Estimate (Standard Error)	Rank	Estimate (Standard Error)	Rank	Estimate (Standard Error)	Rank
Cannabis	2.9439 (0.2141)	5	2.8699 (0.1943)	4	2.7701 (0.1855)	4
Heroin	3.0561 (0.3736)	10	3.4178 (0.3966)	10	3.3160 (0.3871)	10
Cocaine	3.9751 (0.3983)	14	4.2266 (0.4675)	14	4.1827 (0.4882)	14
Magic Mushrooms	3.1977 (0.3255)	11	3.5408 (0.3390)	11	3.4493 (0.3150)	11
Crack	2.9674 (0.3477)	9	3.3103 (0.3587)	9	3.3021 (0.3679)	9
Methadone	2.8412 (0.3219)	7	3.2548 (0.3429)	6	3.1527 (0.3177)	7
Ecstasy	4.8256 (0.5913)	15	4.3968 (0.4819)	15	4.2751 (0.5410)	15
Amphetamines	3.4997 (0.3807)	12	3.9083 (0.4465)	13	3.7310 (0.3862)	13
LSD	3.5474 (0.4196)	13	3.7161 (0.4177)	12	3.6610 (0.4663)	12
Poppers	3.0659 (0.2625)	8	3.2619 (0.2507)	7	3.1494 (0.2648)	6
Ketamine	2.5000 (0.3035)	3	2.9295 (0.3343)	5	2.9520 (0.3412)	5
Anabolic Steroids	2.3787 (0.3010)	2	2.6123 (0.3012)	3	2.5809 (0.3244)	3
Gas	1.0061 (0.0782)	1	1.0426 (0.0761)	1	1.0630 (0.0769)	1
Other drugs	2.5786 (0.3211)	4	2.6027 (0.3056)	2	2.5413 (0.3110)	2
Tranquillisers	2.7815 (0.3718)	6	3.2847 (0.4049)	8	3.2410 (0.3736)	8

Table 6.4.1: Table of Estimates of Discrimination Factor with Two Imputation Schemes and Complete case Analysis in Item Response Theory Model. Notice that the ranks for the parameters using the two imputation models were identical, but a large number of differences between the ranks of the complete case analysis were identified.

Difficulty Factor						
	Complete case		MICE Scheme 1		MICE Scheme 2	
Drug	Estimate (Standard Error)	Rank	Estimate (Standard Error)	Rank	Estimate (Standard Error)	Rank
Cannabis	1.5766 (0.0415)	1	1.5493 (0.0387)	1	1.5562 (0.0397)	1
Heroin	3.0661 (0.1452)	11	2.8364 (0.1160)	11	2.7856 (0.1169)	11
Cocaine	2.5016 (0.0782)	4	2.3962 (0.0721)	3	2.3671 (0.0739)	3
Magic Mushrooms	2.5222 (0.0961)	5	2.4013 (0.0736)	4	2.3864 (0.0704)	4
Crack	3.0180 (0.1448)	10	2.7890 (0.1047)	10	2.7466 (0.1033)	10
Methadone	3.0038 (0.1437)	9	2.7663 (0.1055)	9	2.7284 (0.1011)	9
Ecstasy	2.4781 (0.0800)	3	2.4242 (0.0703)	5	2.4056 (0.0807)	5
Amphetamines	2.7091 (0.1054)	6	2.5455 (0.0879)	6	2.5345 (0.0849)	6
LSD	2.8573 (0.1168)	8	2.7584 (0.1000)	8	2.7216 (0.1128)	8
Poppers	2.3403 (0.0771)	2	2.2605 (0.0597)	2	2.2563 (0.0660)	2
Ketamine	3.2287 (0.1875)	13	2.9291 (0.1396)	12	2.8631 (0.1282)	12
Anabolic Steroids	3.3680 (0.2122)	15	3.1431 (0.1636)	14	3.0763 (0.1897)	14
Gas	2.7954 (0.1704)	7	2.6984 (0.1522)	7	2.6439 (0.1474)	7
Other drugs	3.2609 (0.1891)	14	3.1853 (0.1709)	15	3.1278 (0.1739)	15
Tranquillisers	3.2190 (0.1871)	12	2.9463 (0.1312)	13	2.8774 (0.1158)	13

Table 6.4.2: Table of Estimates of Difficulty Factor with Two Imputation Schemes and Complete case Analysis in Item Response Theory Model. Notice that the ranks for the parameters using the two imputation models were identical, but a large number of differences between the ranks of the complete case analysis were identified.

From Table 6.4.1, the estimates of the discrimination factors of gas were found to be around 1, and the rest of the drug-trying response variables were found to be between 2 and 4, in respect of data sets from each of the two MICE schemes and the complete case analysis. This observation reflected that, except gas, which has an average level of separation, each of the 15 drugs has a high degree of separation and thus exerted an impact on the students' overall drug-trying behaviour. Amongst the 15 drugs, the estimates of the discrimination factors computed under each of the three schemes (i.e. the two MICE schemes and the complete case analysis) of data sets consistently showed that ecstasy, cocaine, amphetamines, LSD, magic mushrooms and heroin were ranked the top six drugs in terms of their high mean estimate values with ecstasy yielding the highest estimates. In other words, the aforesaid six drugs were expected to exert higher influence on the students' drug-trying propensity. On the other hand, anabolic steroids, other drugs and gas were consistently ranked the bottom three drugs in terms of their low mean estimate values with gas yielded the lowest mean estimate value of around 1.

From Tables 6.4.2, the estimates of the difficulty factors of all the 15 drug-trying response variables, computed under each of the three schemes (i.e. the two MICE schemes and the complete case analysis) of data sets, were found to be greater than 1.5, with the majority found to be between 2.5 and 3.2. This observation generally reflected the low proportion of the students who had ever tried each of the 15 drugs. However, amongst the 15 drugs, the estimates of the difficulty factors computed under each of the three schemes (i.e. the two MICE schemes and the complete case analysis) of data sets consistently showed that cannabis, poppers, cocaine, magic mushrooms, ecstasy and amphetamines have relatively lower mean estimate values with cannabis yielded the lowest estimates. This reflected the highest proportion of the students who had tried cannabis, and poppers, cocaine, magic mushrooms, ecstasy and amphetamines.

On the other hand, tranquillisers, anabolic steroids and other drugs were consistently found to have relatively higher estimates with anabolic steroids yielded the highest estimate. This implied that tranquillisers, anabolic steroids and other drugs yielded low proportions of the students who had tried these drugs.

From Table 6.4.1, when examining the standard errors in respect of all the estimates of the discrimination factors, it was observed that for the complete case analysis, the standard error range was between 0.2141 and 0.5913. Also, for MICE scheme 1, the standard error range was between 0.1943 and 0.4819, and for MICE scheme 2, the standard error range was between 0.1855 and 0.5410. These standard error ranges, though differed slightly, did not include extreme values. In terms of order of ability to discriminate, as seen from the "rank" column in Table 6.4.1, the results from the two MICE schemes appeared to be similar.

From Table 6.4.2, when examining the standard errors in respect of all the estimates of the difficulty factors, it was observed that for the complete case analysis, the standard error range was between 0.0415 and 0.2122. Also, for MICE scheme 1, the standard error ranges were between 0.0387 and 0.1709, and for MICE scheme 2, the standard error range was between 0.0397 and 0.1897. Similar to the discrimination factor, these standard error ranges of the estimates of the difficulty factors though differed slightly, they included no extreme values. In terms of order of position of difficulty factor, as seen from the "rank" column in Table 6.4.2, the results from the two MICE schemes appeared to be similar.

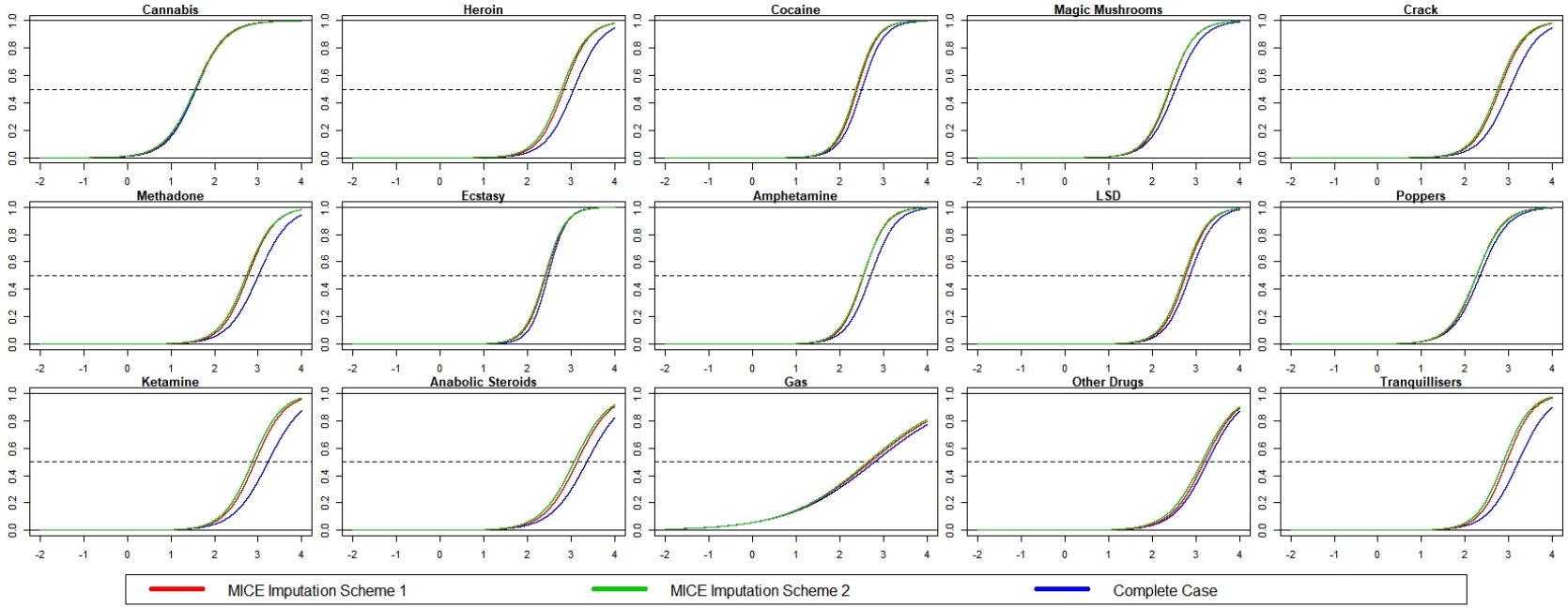

Figure 6.5: Item Characteristic Curves for Two Imputation Schemes and Complete Case Analysis of Item Response Theory Model, for 15 drug-trying response variables. The curves towards the left indicate a higher probability of trying any drug. Red represents MICE Scheme 1, green represents MICE Scheme 2 and blue represents complete case analysis. Curves with steeper slopes indicate better separation among respondents of a scheme for the corresponding drug

Figure 6.5 illustrated the analogous item characteristic curves of the 15 drug-trying response variables in respect of the complete case analysis and the two MICE schemes. For the complete case analysis, the item characteristic curves of heroin, crack, methadone, ketamine, anabolic steroids and tranquillisers were slightly separated from the corresponding curves for the two MICE schemes. On the other hand, the item characteristic curves based on MICE schemes 1 and 2 were all at the leftmost. This comparison implied that the results based on MICE schemes 1 and 2 provided the most optimistic estimation, with a relatively higher possibility of trying any drug by the students. On the contrary, the complete case analysis provided relatively lower possibility estimates of trying any drug by the students at any factor score point. Finally, the item characteristic curve of ecstasy was found to be the steepest, whereas the curve of gas was found to be the flattest. This observation was supported by the findings that ecstasy yielded the highest estimated mean value and gas yielded the lowest estimate mean value in the completed analysis case and the two MICE schemes as shown in Table 6.4.1 in this section.

6.5 Results of Item Response Theory Model under Bayesian Approach

In the Bayesian approach, all the prior combinations mentioned in Table 6.3.1 in Section 6.3.2 were compared for their sensitivity of the item response theory model results to different priors.

Before conducting this sensitivity analysis, we examined the trace plots of the discrimination and difficulty parameters to diagnose their convergence. The related trace plots for the prior combination of discrimination factor prior α_2 and difficulty factor prior δ_1, are presented in Figure 6.6. To discuss the results of the

sensitivity analysis under the Bayesian approach, the tables of posterior means and posterior standard deviations and rankings of each α and δ parameter for each considered prior specifications of α and δ are presented in Appendix D. The results for the prior combination of discrimination factor α_2 and difficulty factor δ_1, as well as prior combination of discrimination factor α_3 and difficulty factor δ_1, are presented in Table 6.5.1. The results of two discrimination factor priors (i.e. α_2 and α_3) are selected to discuss the results of the sensitivity analysis of item response theory models. The estimates of discrimination factor priors α_2 and α_3 are displayed in Table 6.5.1. The plots of the combined 95% confidence intervals in respect of all combinations of priors are presented in Figures 6.7 to 6.10. Their respective item characteristic curves plots are presented in Figures 6.11 and 6.12.

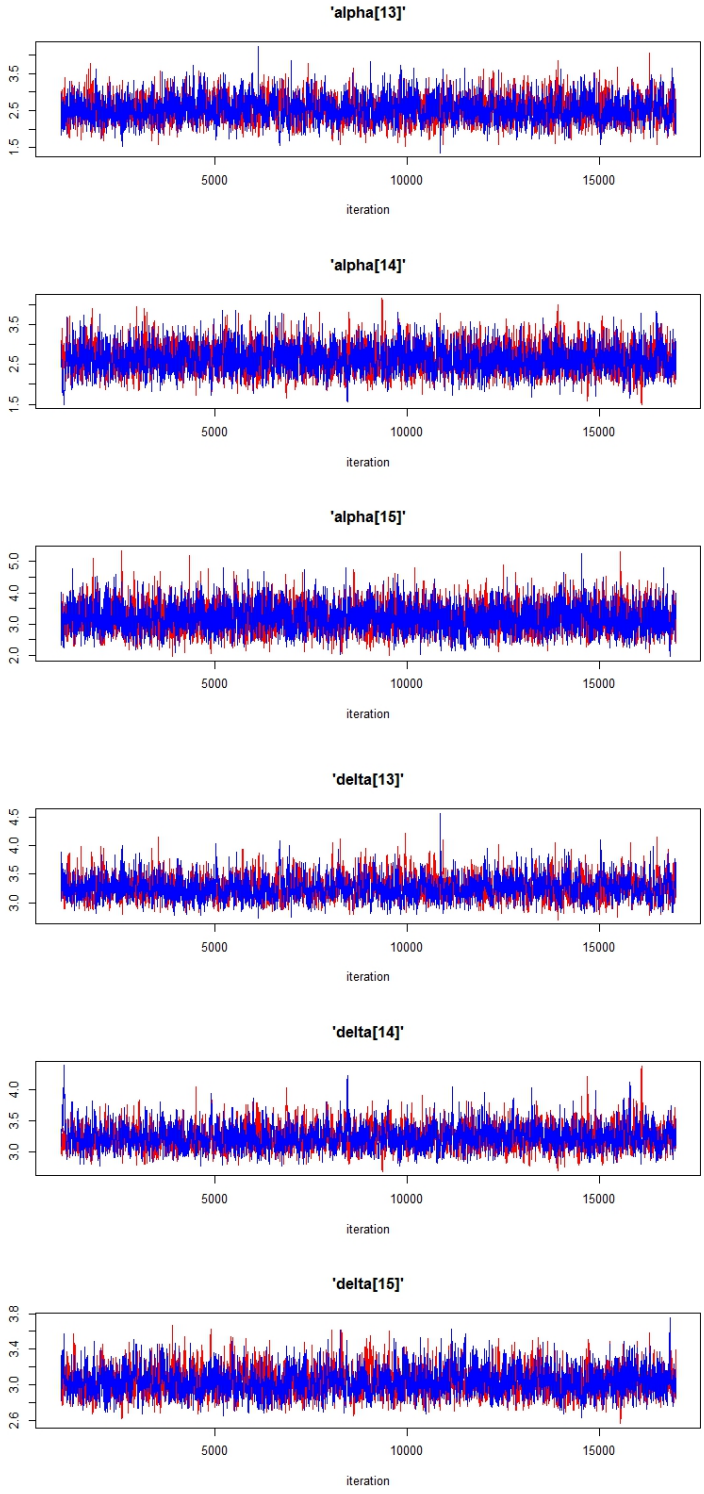

Figure 6.6: Trace Plots of the Estimates of the Discriminatory Factor Prior α_2 and the Difficulty Factor Prior δ_1. The top three rows depict the posterior mean of the discriminatory factors (α) for Gas, Other Drugs and Tranquillisers response variables respectively, whereas the bottom three rows represent the estimates of the difficulty factors (δ) for Gas, Other Drugs and Tranquillisers response variables respectively.

Table 6.5.1: Table of Posterior Means and Standard Deviations of Discrimination and Difficulty Factors with Discrimination Priors α_2 and α_3. For details about prior distributions, please refer to Table 6.3.1.

		α_2		α_3	
		Posterior mean(sd)	Rank	Posterior mean(sd)	Rank
Discrimination Factor					
δ_1	Cannabis	2.853 (0.1969)	4	2.865 (0.205)	5
	Heroin	3.434 (0.4061)	11	3.342 (0.3951)	11
	Cocaine	4.337 (0.4393)	14	4.198 (0.47)	14
	Magic Mushrooms	3.391 (0.3033)	10	3.31 (0.2882)	10
	Crack	3.323 (0.3714)	9	3.211 (0.3565)	9
	Methadone	3.199 (0.3604)	7	3.098 (0.3355)	7
	Ecstasy	4.836 (0.5954)	15	4.576 (0.5055)	15
	Amphetamines	3.935 (0.4028)	12	3.849 (0.4004)	13
	LSD	3.946 (0.4925)	13	3.742 (0.4282)	12
	Poppers	3.244 (0.2443)	8	3.184 (0.2275)	8
	Ketamine	2.921 (0.3153)	5	2.812 (0.284)	4
	Anabolic Steroids	2.531 (0.2709)	2	2.433 (0.283)	2
	Gas	1.014 (0.07113)	1	0.9974 (0.06987)	1
	Other Drugs	2.617 (0.3226)	3	2.547 (0.2898)	3
	Tranquillisers	3.167 (0.4117)	6	3.022 (0.3641)	6
Difficulty Factor					
δ_1	Cannabis	1.569 (0.03857)	1	1.568 (0.04119)	1
	Heroin	2.902 (0.1169)	11	2.936 (0.1178)	11
	Cocaine	2.433 (0.06683)	3	2.454 (0.06884)	3
	Magic Mushrooms	2.466 (0.07425)	5	2.481 (0.07537)	5
	Crack	2.845 (0.1115)	10	2.884 (0.1151)	10
	Methadone	2.828 (0.1148)	9	2.86 (0.1095)	9
	Ecstasy	2.436 (0.06398)	4	2.468 (0.06733)	4
	Amphetamines	2.582 (0.0767)	6	2.606 (0.0805)	6
	LSD	2.765 (0.09819)	7	2.806 (0.09764)	7
	Poppers	2.294 (0.06131)	2	2.314 (0.05906)	2
	Ketamine	2.98 (0.1392)	12	3.029 (0.1316)	12
	Anabolic Steroids	3.239 (0.1699)	15	3.313 (0.1855)	15
	Gas	2.78 (0.1558)	8	2.812 (0.1527)	8
	Other Drugs	3.227 (0.1679)	14	3.273 (0.1805)	14
	Tranquillisers	3.029 (0.142)	13	3.09 (0.1559)	13

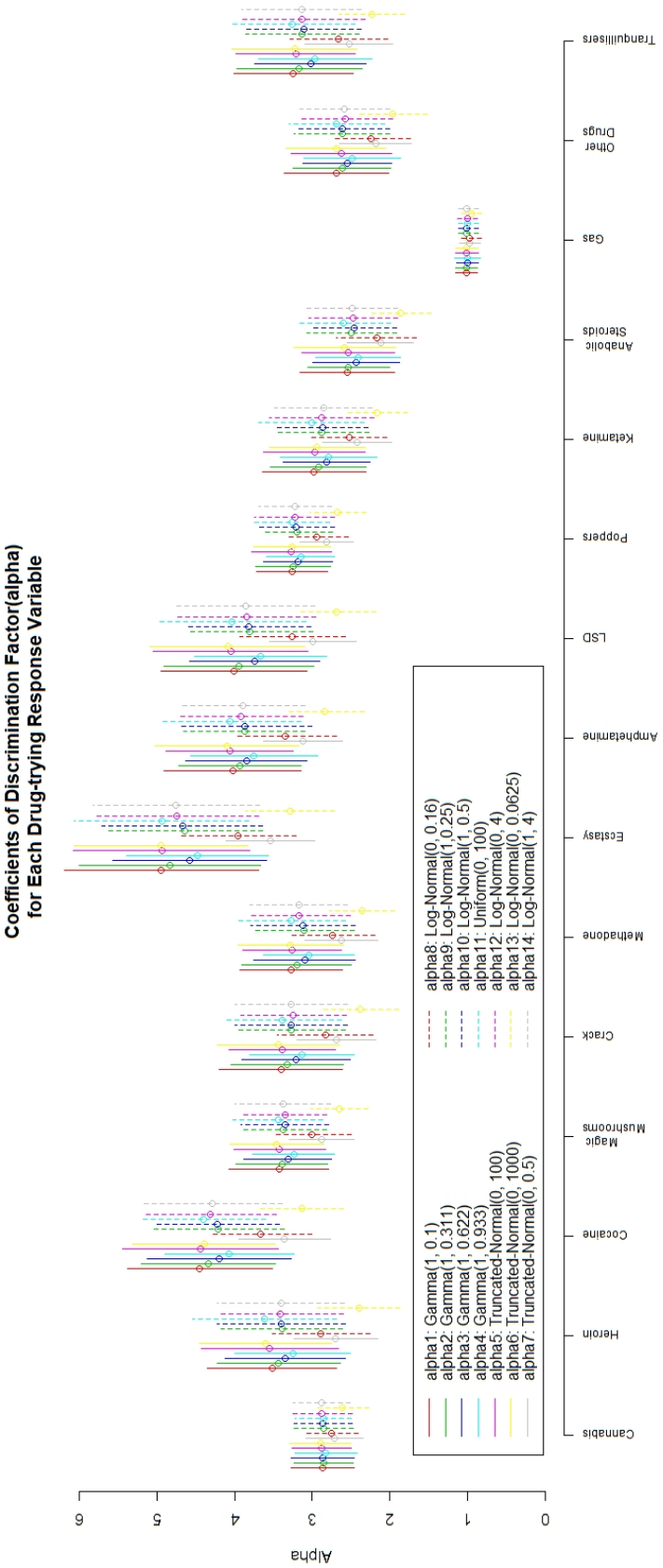

Figure 6.7: Confidence Interval Plots of Discrimination Factor for IRT Model in OpenBUGS with δ_1 Prior. Except for the models with truncated-normal(0, 0.5) and log-normal(0,0.0625) priors, the mean estimates of the IRT models with different priors are similar. Note the consistent results for gas.

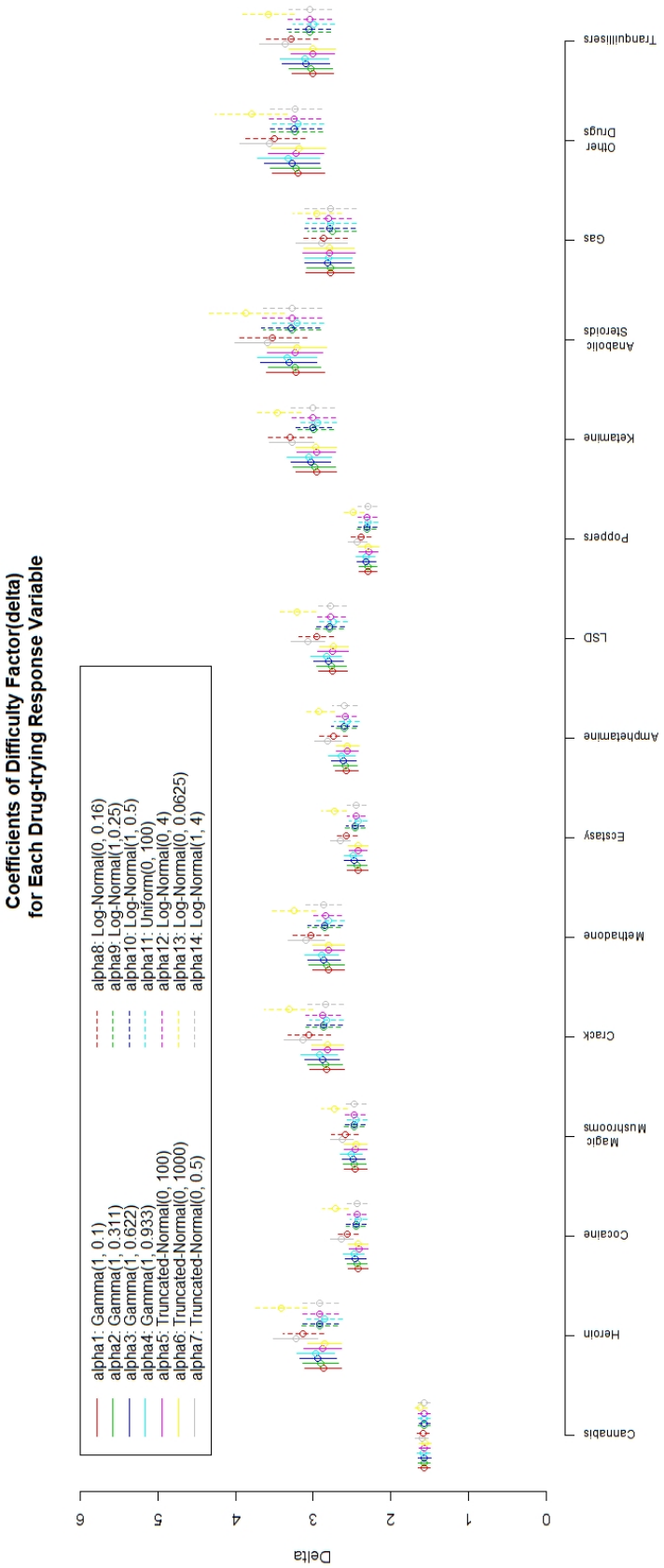

Figure 6.8: Confidence Interval Plots of Difficulty Factor for IRT Model in OpenBUGS with δ_i Prior. Except for the models with truncated-normal(0, 0.5) and log-normal(0,0.0625) priors, the mean estimates of the IRT models with different priors are similar. Note the consistent results for cannabis.

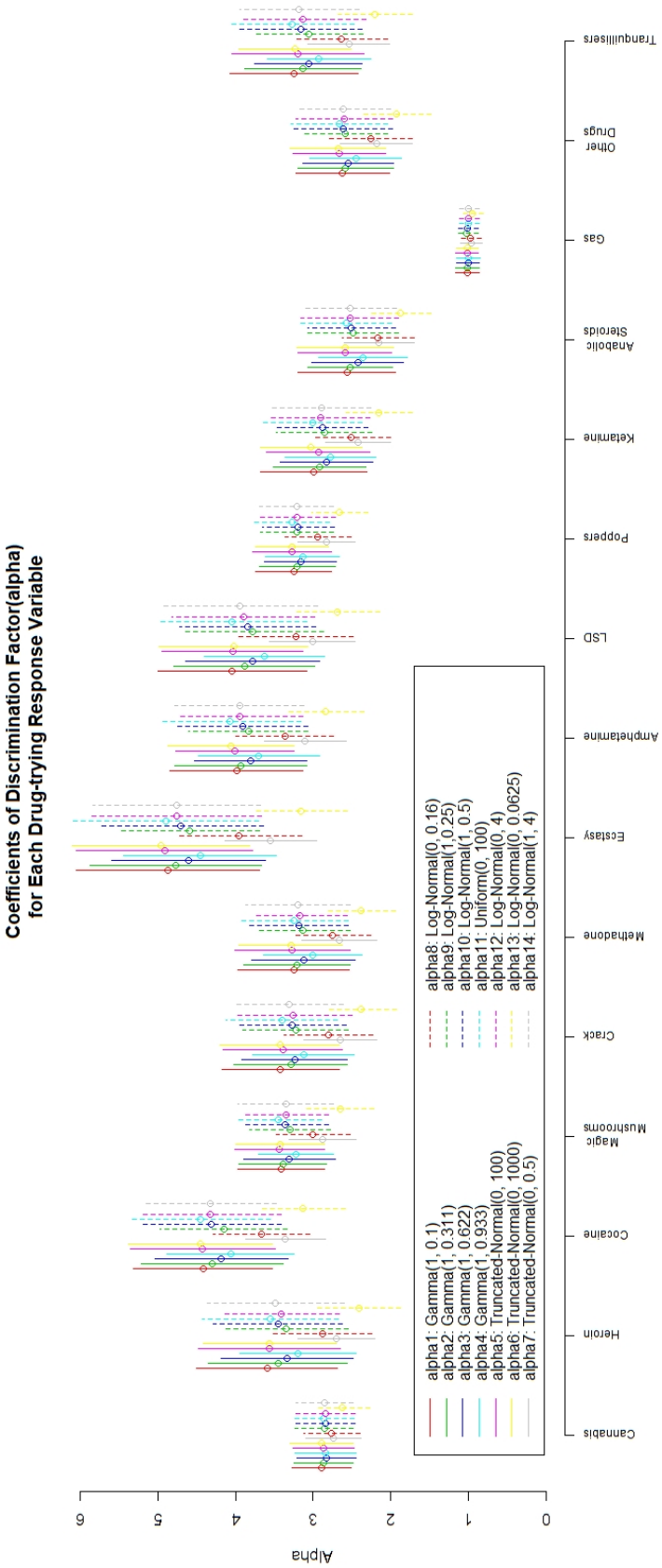

Figure 6.9: Confidence Interval Plots of Discrimination Factor for IRT Model in OpenBUGS with δ_2 Prior. Except for the models with truncated-normal(0, 0.5) and log-normal(0,0.0625) priors, the mean estimates of the IRT models with different priors are similar. Note the consistent results for gas.

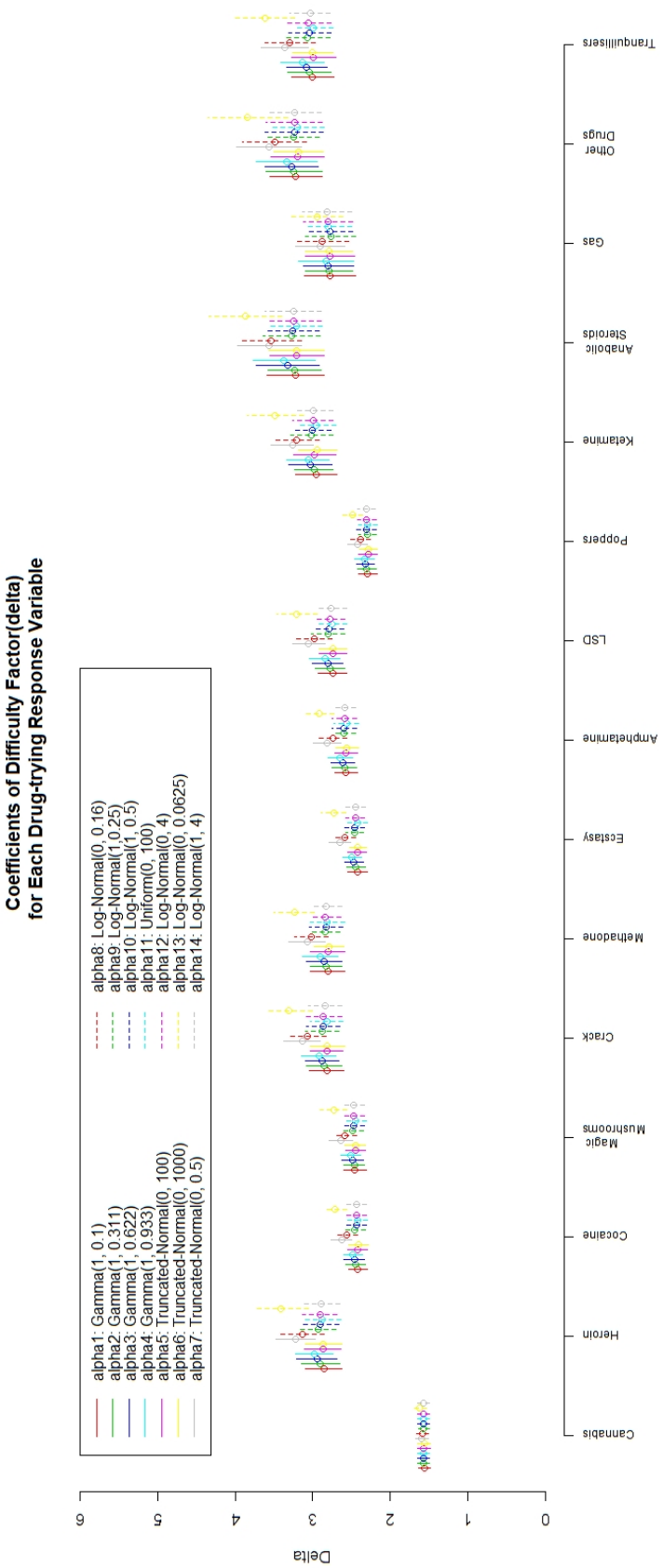

Figure 6.10: Confidence Interval Plots of Difficulty Factor for IRT Model in OpenBUGS with δ_2 Prior. Except for the models with truncated-normal(0, 0.5) and log-normal(0,0.0625) priors, the mean estimates of the IRT models with different priors are similar. Note the consistent results for cannabis.

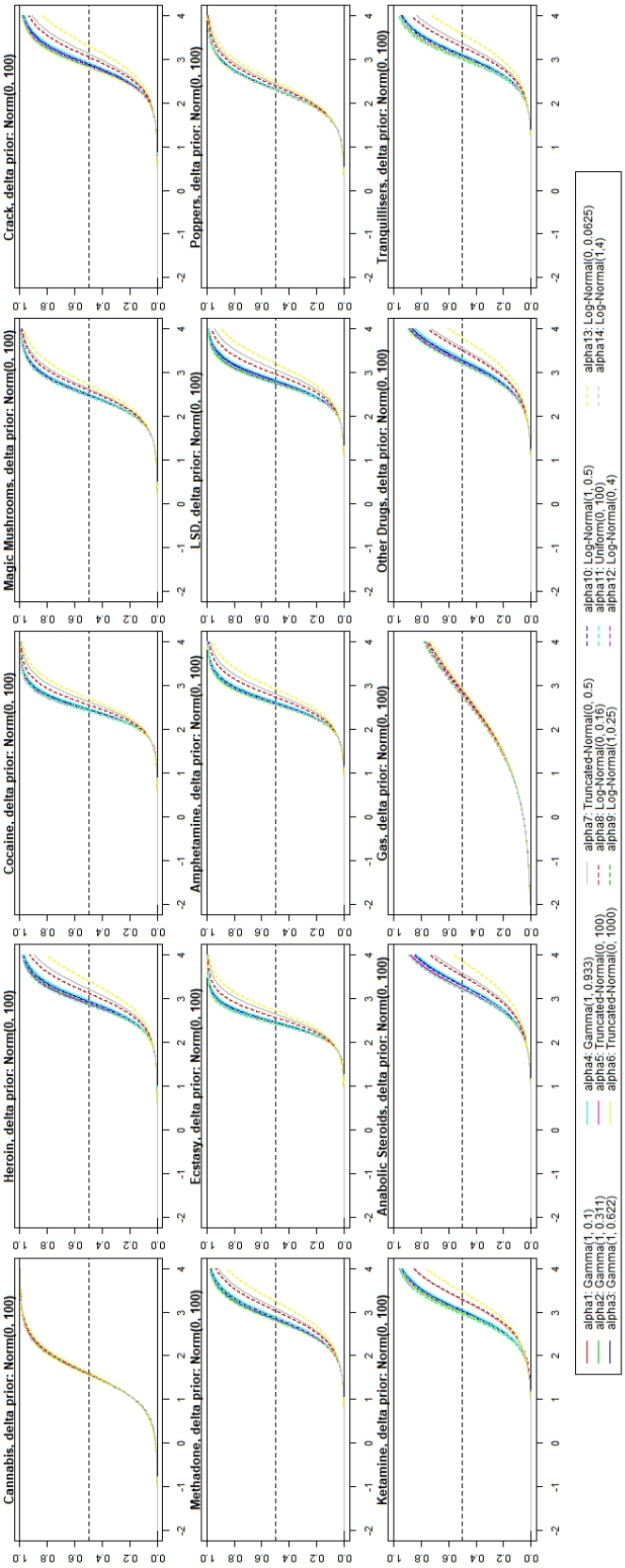

Figure 6.11: Item Characteristic Curves for IRT Model in OpenBUGS with δ_1 Prior

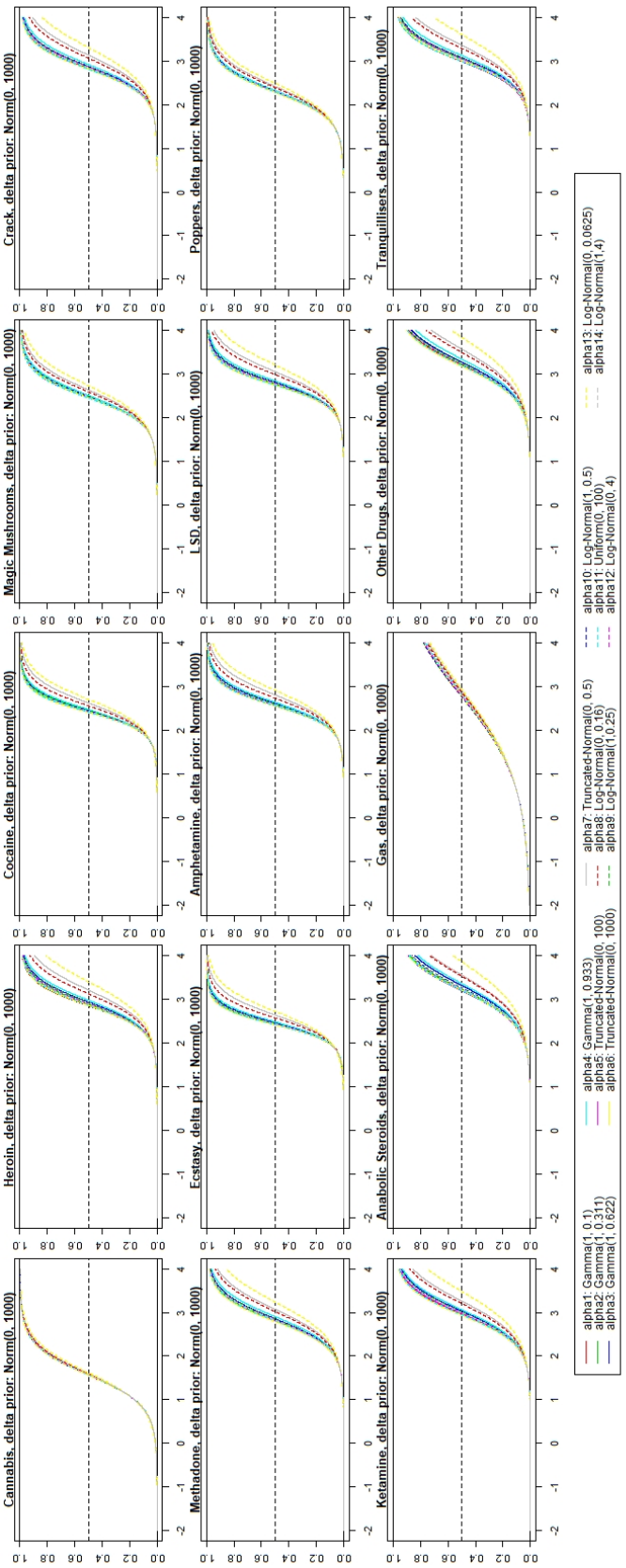

Figure 6.12: Item Characteristic Curves for IRT Model in OpenBUGS with δ_2 Prior

As shown in Table 6.5.1, the estimates of the discrimination factors for gas was found to be 1, and the estimates of the discrimination factors for the remaining 14 drug-trying response variables were found to be between 2 and 4. Amongst the 15 drugs, the estimates of the discrimination factors showed that ecstasy, cocaine, amphetamines, LSD, heroin and magic mushrooms were ranked the highest six drugs, in terms of their high posterior means with ecstasy yielded the highest posterior mean. On the other hand, other drugs, anabolic steroids and gas were ranked the bottom three drugs in terms of their low posterior means, with gas yielded the lowest posterior mean of around 1. The posterior means of the difficulty factors of all the 15 drug-trying response variables were found to be greater than 1.5, with the majority found to be between 2.5 and 3.2. Also, amongst the 15 drugs, the posterior means of the difficulty factors showed that cannabis, poppers, cocaine, ecstasy, magic mushrooms and amphetamines have relatively lower mean estimate values with cannabis yielded the lowest posterior mean. On the other hand, tranquillisers, other drugs and anabolic steroids were found to have relatively higher posterior mean with anabolic steroids yielded the highest posterior mean. These aforesaid findings in respect of the estimates of the discrimination factors and difficulty factors of the 15 drug-trying response variables were found to be consistent with the corresponding findings under the marginal approach as discussed in Section 6.4 above.

From Figures 6.8 and 6.10, it was observed that for both difficulty factor priors, δ_1 and δ_2, the estimates and the orderings of difficulty factor generated from all the discrimination factor priors were similar, except for those in respect of discrimination factor priors α_7, α_8 and α_{13}, which were slightly different from other posterior means. Figures 6.7 and 6.9 illustrated slight downward distortion in discrimination factor estimates for priors α_7, α_8 and α_{13}, and slight upward distortion in difficulty factor estimates for priors α_7, α_8 and α_{13} were shown

in Figures 6.8 and 6.10. Another observation was that the standard deviations were smaller when the estimates were smaller, leading to a higher degree of precision. The discrepancies of standard deviations for the discrimination factor among discrimination factor priors might explain the increasing certainty of drug responses which have less separation power of the students who had tried drugs from those who had not. Similarly, the discrepancies of standard deviations for the difficulty factor among discrimination factor priors might explain the increasing certainty of drug responses which yield larger proportions of the students who had tried them.

Furthermore, it was observed from Figures 6.11 and 6.12 that the item characteristic curves in respect of difficulty factor priors δ_1 and δ_2 were similar, except for the curves representing the two MICE schemes and the complete case analysis respectively which were packed relatively denser in plots related to difficulty factor prior δ_2. From Tables D.1.1 to D.1.6 in Appendix D, the rankings of the estimates in respect of discrimination factor priors were similar, except for discrimination factor priors α_7, α_8 and α_{13}. The discrepancies between the rankings might due to close estimates, where their differences were within a small fraction of standard errors. In general, it could be concluded that the estimates of the discrimination and difficulty factors were insensitive to discrimination factor priors and difficulty factor priors.

In short, the Figures 6.7 to 6.10 summarised the general observation that most discrimination factor priors generated similar results on discrimination factor and difficulty factor estimates. The imputed data sets from the two MICE schemes under Bayesian approach generated similar results.

In Section 6.6, a comparison between the results generated in the `R` program and those generated in the `OpenBUGS` program is made and discussed.

6.6 Comparison between Marginal Approach and Bayesian Approach and Limitation

In this final section, we firstly discuss the differences in methods for handling missing data between the `OpenBUGS` program and the `R` program, as well as the advantages and disadvantages of both programs. We then compare the estimates and standard errors of the item response theory models from the `R` program with those from the `OpenBUGS` program to investigate the extent of their discrepancies. The results from the data sets imputed only through the MICE imputation with the 15 drug-trying response variables (i.e. MICE scheme 1) were chosen to represent the `R` program, whereas the results generated under the discrimination factor prior α_2 and the difficulty factor prior δ_1 were selected to represent the results from the `OpenBUGS` program. At the end of this section the advantages and drawbacks of using the item response model were discussed.

In the marginal approach, there were different procedures of imputing the data with the following two configurations: (1) MICE imputation with 15 drug-trying response variables only (i.e. MICE scheme 1) and (2) MICE imputation with 15 drug-trying response variables and covariates (i.e. MICE scheme 2). Before the marginal approach was implemented, the missing data were firstly imputed through the "MICE" imputation, in which the missing values were imputed for each variable at a time, conditioning on the rest of the variables as covariates in the data set with pre-defined distributions to generate imputed data sets. In some occasions, the distribution used for imputing missing values might be different from that used for statistical analysis. On the other hand, in the Bayesian approach, the missing data was imputed and updated within the same model for statistical analysis conditioning on the rest of the variables. Basically, the marginal approach was a two-stage approach, whereas the Bayesian approach was a one-stage approach. The marginal approach involved ten imputed data

sets, where their estimates were combined through Rubin's rule. On the other hand, the Bayesian approach involved a single data set with missing values being imputed through at least two chains. Statistical inference of the Bayesian approach involved any single chain instead of all chains.

Furthermore, in the Bayesian approach, initial values for missing data were generated from a random seed, whereas in the marginal approach, the initial values for missing data were the mean of the variable across observed data. However, in the marginal approach, since the parameters of the item response models were based on the imputed data sets, no priors were required for these parameters, whereas in the Bayesian approach, a proper prior was required for every parameter. If an incorrect prior was used, the result might be distorted. Also, both the Bayesian approach and the Marginal approach took into account imputation uncertainty in two different ways.

The estimates and standard error comparison tables of the `R` and `OpenBUGS` programs are generated in Table D.2.1 in Appendix D. The plots of the combined 95% confidence intervals from the two programs are presented in Figures 6.13 and 6.14, followed by their respective item characteristic curve plots in Figure 6.15.

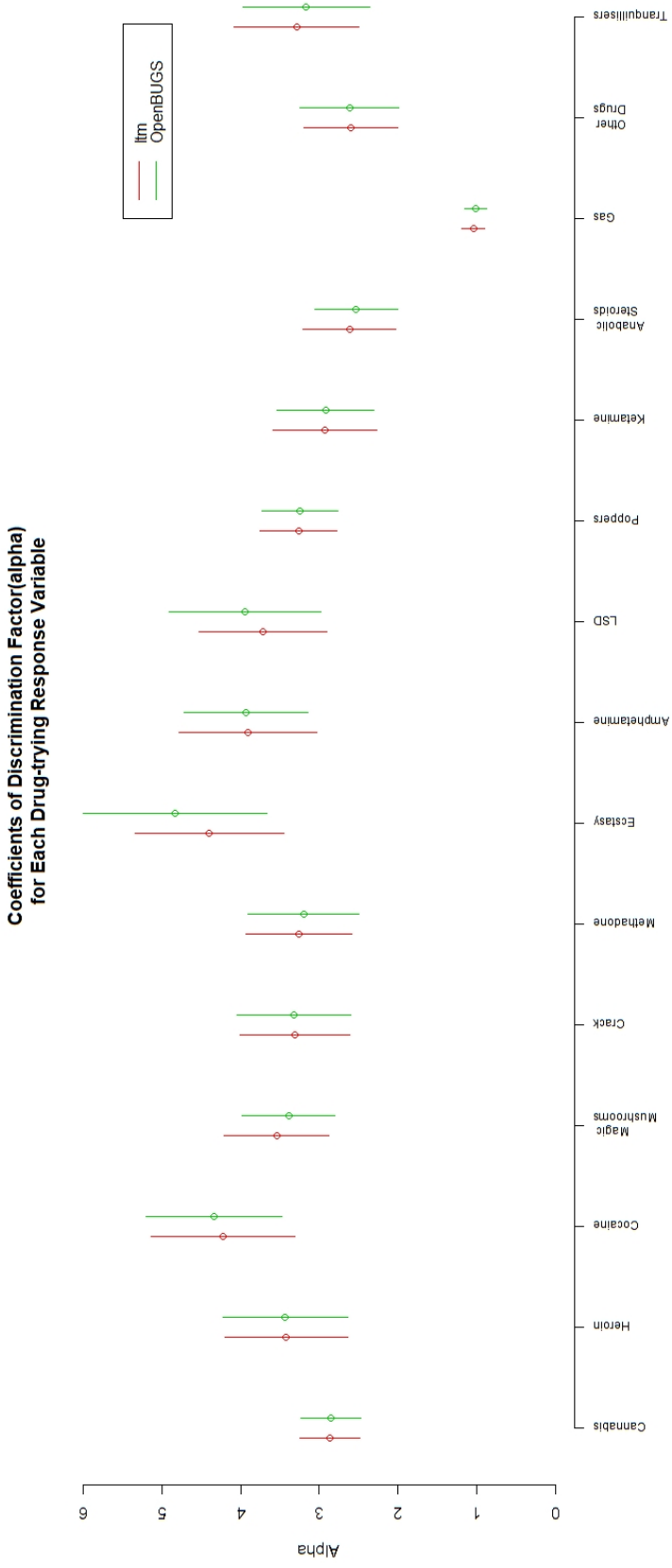

Figure 6.13: Confidence Interval Plots of Discrimination Factor from Two Programs of Item Response Theory Model. The estimates generated from the two programs are similar.

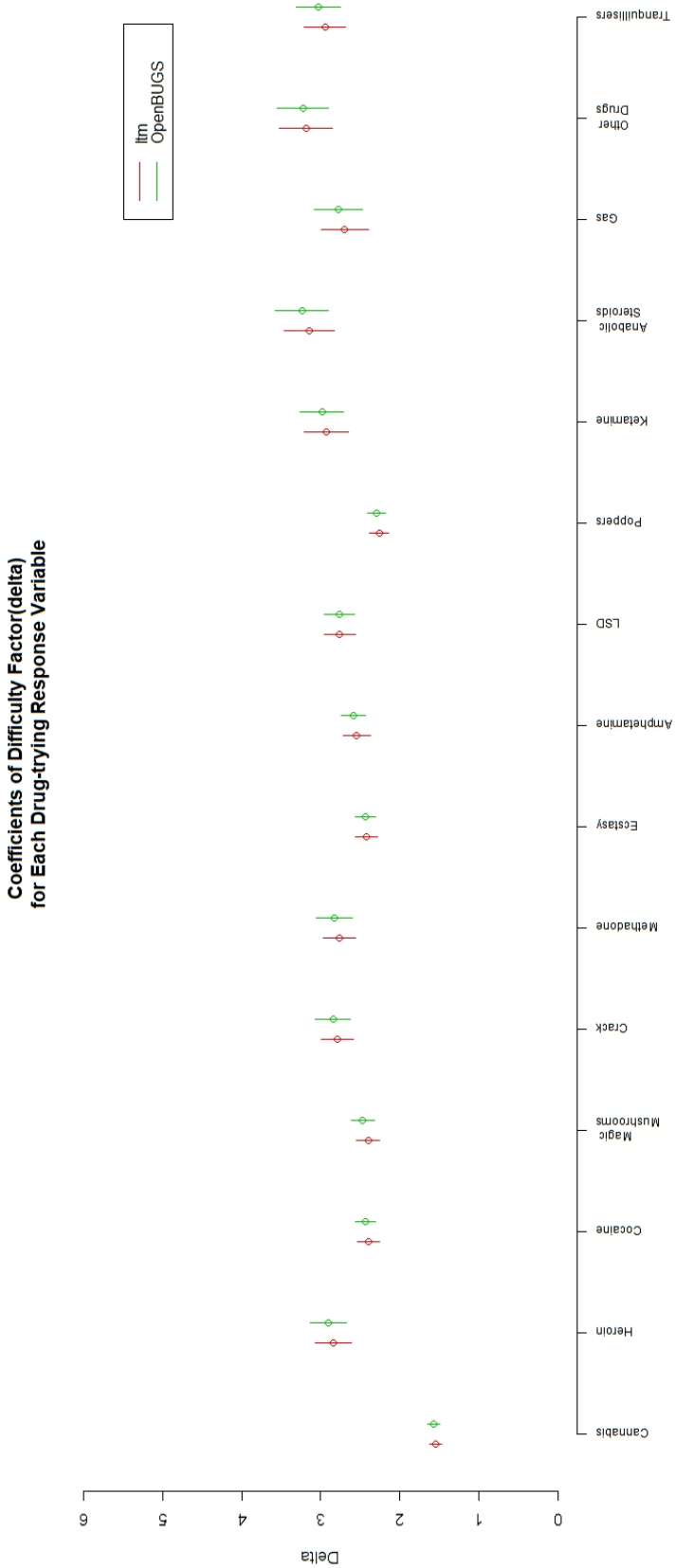

Figure 6.14: Confidence Interval Plots of Difficulty Factor from Two Programs of Item Response Theory Model. The estimates from the `ltm` package in R program are slightly lower than those from the OpenBUGS program.

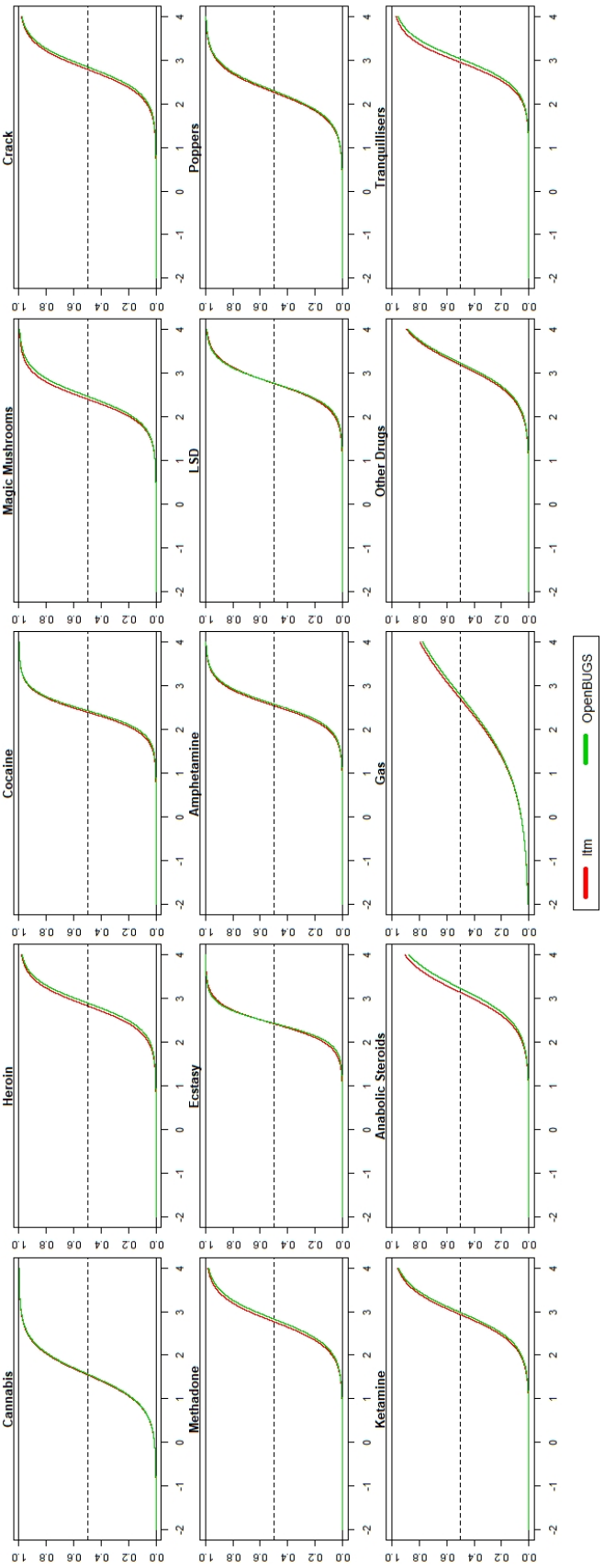

Figure 6.15: Item Characteristic Curves from Two Programs of Item Response Theory Model. Red represents `ltm` and green represents `OpenBUGS`. The plots from the two programs are similar.

As could be seen from Figure 6.13, the plots for the estimates of the discrimination factor generated from the R and OpenBUGS programs appeared to be similar. On the other hand, from Figure 6.14, the plots for each estimate of the difficulty factor generated from the R program showed slightly downward shift when compared to those produced by the OpenBUGS program. From Table D.2.1 in Appendix D, the estimates of the discrimination factor for all the 15 drugs and their respective rankings generated from the R program, were similar to those generated from the OpenBUGS program. Although all the estimates of the difficulty factor for all the 15 drugs generated from the R program were slightly lower than those generated from the OpenBUGS program, their respective rankings were similar. The relatively lower mean estimates of the difficulty factor, which explained the phenomenon of slightly shifting downward of the plots for the estimates, for all the 15 drugs generated by the R program might due to the difference in the algorithms of prediction models for the missing values during imputation.

Finally, from the item characteristic curves in Figure 6.15, we observed that the two curves in every plot were contiguous with each other, with the curves representing the R program on the left side. This observation was consistent with the result tables in Appendix D, along with Figures 6.13 and 6.14.

There are advantages of using the item response theory model in this study, which include: (1) the Item response theory model is good for a data set where a core of items, such as 15 drug-trying response variables in this study, is analysed (Baker, 2001) and (2) the item response theory model helps to describe in a more comprehensive way the associations among the 15 drug-trying response variables by the discrimination and difficulty parameters. Nevertheless, it is worth to note that there are a few drawbacks of using the item response theory model, namely: (1) assumptions of the item response theory model are strong

and (2) estimates generated by the item response theory model are sensitive to variation in sample size (Baker, 2001).

6.7 Summary

In this chapter, the two-parameter item response theory model was implemented on the working data set through two approaches, namely the marginal approach and the Bayesian approach, in order to further investigate the relationships between drug-trying response variables and the students' drug-trying behaviour.

In all the two-parameter item response theory models under marginal approach and Bayesian approach, the estimates of the discrimination factors consistently showed that ecstasy, cocaine, amphetamines, LSD, heroin and magic mushrooms were ranked the top six drugs in terms of their high mean estimate values with ecstasy yielded the highest mean estimate value. On the other hand, other drugs, anabolic steroids and gas were consistently ranked as the bottom three drugs in terms of their low mean estimate values with gas yielded the lowest mean estimate value of around 1. The aforesaid findings shed additional light on the relationships between drug-trying response variables and the students' drug-trying behaviour. Six drugs, namely ecstasy, cocaine, amphetamines, LSD, heroin and magic mushrooms, were found to exert higher influence on the students' drug- trying behaviour that for example, if a student has tried ecstasy, there was a higher likelihood that the student will try other types of drug.

Also, in all the two-parameter item response theory models under marginal approach and Bayesian approach, the estimates of the difficulty factors of all the fifteen drug-trying response variables were found to be greater than 1.5, with the majority found to be between 2.5 and 3.2. This observation generally reflected the low proportion of the students who had ever tried each of the 15 drugs.

However, amongst the 15 drugs, the estimates of the difficulty factors consistently showed that cannabis, poppers, cocaine, magic mushrooms, ecstasy and amphetamines have relative lower mean estimate values with cannabis yielded the lowest mean estimate value. This reflected the highest proportion of the students who had tried cannabis. On the other hand, tranquillisers, anabolic steroids and other drugs were consistently found to have relatively higher mean estimate values with anabolic steroids yielded the highest mean estimate value. The aforesaid findings were consistent with the results shown in a frequency table of Drug-trying Response Variable in Chapter 3.

Estimates of the difficulty and discrimination factors of the two-parameter item response theory models were found to be similar in both the marginal approach and Bayesian approach, albeit the result generated from the `OpenBUGS` program under the Bayesian approach yielded slightly higher difficulty factor estimates across all the fifteen drugs than the result generated from the `R` program under the marginal approach. Such phenomenon might due to the difference in the algorithms of prediction models for the missing values during imputation under marginal approach. All priors, except for a half-normal prior and two log-normal priors for discrimination factor, produced similar results. Such finding supported that estimates were largely non-sensitive in prior changes.

Finally, the Bayesian approach is a slower method than the marginal approach in respect of the item response theory model. Since both approaches generated similar results in this research, deciding on which method is better for employment of the item response theory model became less essential.

In this chapter, we have discussed the item response theory models, where the tendency to try drugs was represented by a continuous latent variable known as a factor score. In Chapter 7, rather than adopting a continuous latent variable,

we investigate the use of drugs by employing a discrete latent variable, which provides clustering information of respondents.

Chapter 7

Latent Class Analysis and K-means Clustering

7.1 Introduction

This chapter outlines the underpinning theory and presents the results of a latent class analysis and a K-means clustering of the working data set.

The univariate logistic regression models in Section 5.2 were adopted to assess the associations between the smoking, drinking and considered drug-related socio-demographic factors and the drug-trying response variables, as well as to assess whether the drug-trying response variables predicted each other. Similarly, the log-linear analysis models, in Section 5.3, aimed to investigate the interactions between the drug-trying response variables, so as to provide insight regarding associations among 15 drugs. The item response theory models in Chapter 6 provided a different perspective of investigating the drug-trying behaviours among students. It permitted an investigation of the proportion of students trying each drug, in terms of the amount of influence of trying each drug on the overall drug-trying behaviours, and the propensity for students to try drugs. However, none of these models sought to classify or cluster students

with respect to their drug-trying behaviour patterns, which may provide additional understanding of any latent sub-structure.

Of interest in this study are the patterns and clusters of the drug-trying behaviours among students, and how these patterns may relate to other covariates. For example, how smoking, drinking and considered drug-related socio-demographic variables are related to any classification of students. At the first stage, we thus need to find criteria by which to classify the students in order to gain insight into their drug-trying behaviour within each classification. To classify the students, we employ the statistical technique known as "cluster analysis" that groups the students' drug-trying behaviour into classes. A brief overview of cluster analysis is presented below.

The term "cluster analysis" was firstly coined by Edwards and Cavalli-Sforza (1965) as identifying "clusters of points in space". "Cluster analysis" includes, but is not limited to, two statistical approaches, namely latent class analysis (Everitt et al., 1993) and K-means clustering (Hartigan, 1975). In this research, we apply both the latent class analysis and K-means clustering, for which we introduce new methodology to enable modelling over multiply imputed data.

In a latent class analysis, the classifying criteria are represented by a latent discrete variable, which classifies respondents into groups (Collins and Lanza, 2010). Generally, a latent class analysis models the patterns of categorical responses and classifies respondents into a specified number of groups via a latent discrete variable (Collins and Lanza, 2010). Latent class analysis has often been used in analysing biological or social data. For example, Agrawal (2006) applied latent class analysis to drug abuse data, in order to characterise poly-substance abuse dependence of respondents of the National Epidemiological Survey on Alcohol and Related Conditions in America. Also, Pharris (2011) applied the

latent class analysis to investigate the relationship between HIV infection and HIV-related risk factors, including drug use, via a latent variable with three classes, which classified respondents by HIV-stigma. The alternative analyses of latent class analysis could be principle component analysis and factor analysis. However, we employed latent class analysis and considered principle component analysis and factor analysis not suitable in this study due to the following reasons: (1) factor analysis can only be used on continuous variables (Hair et al., 1994) and (2) principle component analysis can only be used on variables which follow normal distribution (Bartholomew et al., 2011). Also, interpretation of factor analysis requires an extensive effort and is based on heuristic approach, which might not be a complete method (Hair et al., 1994).

In this research, we apply latent class analysis to examine the potential partitioning of the students in this survey into a specific number of classes, based upon their drug-trying patterns, as well as the proportion of the students in each class and the 'class-conditional' proportions of the students trying each drug. Through this application, we identify the optimal number of classes by maximum likelihood solutions, i.e. AIC and BIC, which adequately explains any latent sub-structure. In addition, latent class analysis can be combined with a logistic regression model to form a latent class regression model, to explain the relationship between class membership and the smoking, drinking and considered drug-related socio-demographic factors via a regession model on a latent variable. In other words, the latent class analysis provides insight by fitting the working data set using a model that partitions the students into some classes based upon their drug-trying patterns, while linking covariates to class membership. A latent class analysis may thus assist in understanding patterns of behaviours and also explaining the relationships between the smoking, drinking and considered drug-related socio-demographic covariates, as well as class membership, and thus enable investigation of the drug-trying behaviour

of young people in greater depth.

In contrast, K-means clustering is a distance-based algorithm which classifies the students into k clusters by minimising the total squared error distance between each student and the cluster mean point within response variables for each corresponding cluster (Jain et al., 1999). The K-means clustering has provided a simple and widely used clustering algorithm for over 50 years (Jain, 2010). The K-means clustering has been applied to various fields, such as medical (Ng et al., 2006) and environmental data (Shi and Zeng, 2014). Figure 7.1 provides a visualisation of three centroids in K-means clustering.

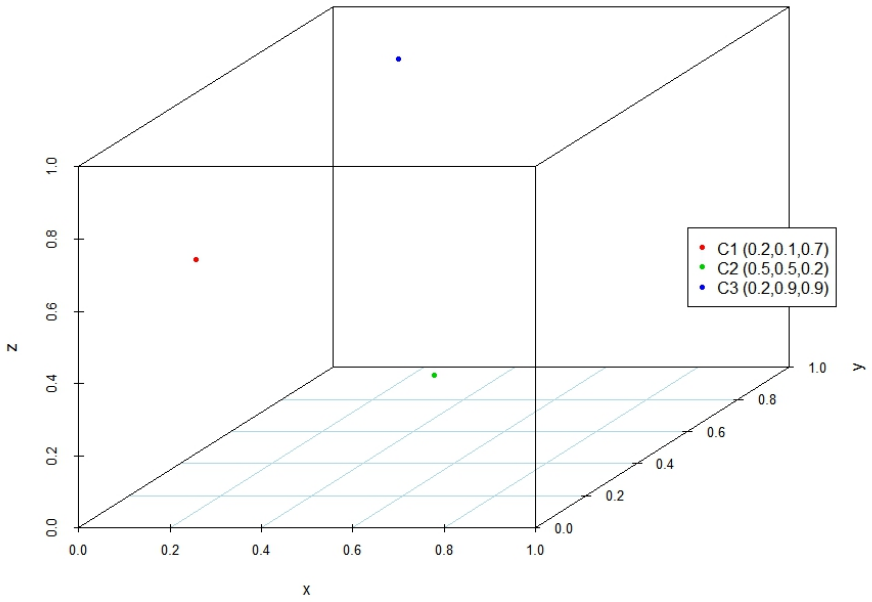

Figure 7.1: Three-dimensional Visualisation of K-means Clustering with Three Centroids, C_1, C_2 and C_3

Through identifying a parsimonious number of clusters and hereafter grouping the students according to their nearest centroid, with respect to the hypercube distance between K drug-trying response variables, the K-means clustering is implemented in this study to provide another perspective on classification of the

students. The results of the K-means clustering can then be compared with that of the latent class analysis to check the validity of classification of the students by the latter.

In the rest of this chapter, Section 7.2 discusses the latent class analysis. Section 7.3 presents and discusses the K-means clustering, which is compared to the results generated from the latent class analysis in Section 7.2.

7.2 Latent Class Analysis

In this section, we firstly introduce latent class analysis contextually, before presenting the theory and application of the latent class analysis to the working data set. We conclude with a discussion of the results and utility of the latent class analysis.

7.2.1 Introduction

Latent class analysis was firstly proposed by Lazarsfeld (1950) as measuring models for categorical response data. It was until Goodman (1974), who proposed the latent class analysis under the name "latent structure analysis", proposed a maximum likelihood procedure for estimating the latent class analysis models. Dayton and MacReady (1988) introduced a type of latent class analysis, under the name "latent class models", in which the probabilities of latent class memberships were functionally and directly related to concomitant variables. The name "latent class analysis" was coined by Everitt et al. (1993).

In this research, we use the latent class analysis to classify the students' drug-trying patterns with a discrete latent variable, which discerns the number of classes required to classify the students. We also implement a latent class regression model to establish the linkage between the drug-trying response vari-

ables and smoking, drinking and considered drug-related socio-demographic predictors via the discrete latent variable.

We also implement the methodology of selection of the optimal number of classes, as well as the backward elimination by Rubin's rule with Wald's test in the latent class analysis.

7.2.2 Theory

Following the exposition in Dayton and MacReady (1988), the proportions of latent classes, the class-conditional probabilities, the likelihood, the marginal distribution and the constraints of latent class analysis model are defined in this section. Suppose that there are L latent classes, n individuals, and J' categorical variables. Also, for each i^{th} individual, $i = 1, \ldots, n$, each j'^{th} categorical variable, $j' = 1, \ldots, J'$ contains $R_{j'}$ response levels. Let $Y = \{y_1, \ldots, y_n\}$ be the data set matrix, where $y_i = \{y_{i1}, \ldots, y_{iJ'}\}$ for each vector of data set points for each individual i and for each j'^{th} categorical variable, $j' = 1, \ldots, J'$. For each response level r of each categorical variable j', $r = 1, \ldots, R_{j'}$ and each latent class $l = 1, \ldots, L$, let p_l be a proportion of latent class l, and let $\pi_{j'rl}$ be the class-conditional probability of response level r to variable j' in latent class l. Then the likelihood of the general latent class analysis model for each latent class c_l, in a collection of the latent class analysis models, $C(l = 1, \ldots, L)$, (i.e. $c_l \in C(l = 1, \ldots, L)$), is defined as:

$$f(Y \mid c_l) = \prod_{j'=1}^{J'} \prod_{r=1}^{R_{j'}} \pi_{j'rl}^{\delta_{ij'}}, \tag{7.1}$$

and the marginal distribution of the general latent class analysis model is defined as:

$$f(Y) = \sum_{l=1}^{L} p_l f(Y \mid c_l), \tag{7.2}$$

where the $\delta_{ij'}$ is the Kronecker Delta, which acts as an indicator of whether the option was chosen. The Kronecker Delta is defined as:

$$\delta_{ij'} = \begin{cases} 1, & y_{ij'} = r \\ 0, & \text{otherwise.} \end{cases} \tag{7.3}$$

Equation 7.3 explains that the latent class analysis model uses a dummy-variable system, which identifies whether the response level r is taken for categorical variable j' by individual i. The probability parameters are subjected to the following restraints - all class conditional probabilities within all responses of categorical variable j' are summed up to 1:

$$\sum_{r=1}^{R_{j'}} \pi_{j'rl} = 1, j' = 1, \ldots, J', l = 1, \ldots, L. \tag{7.4}$$

In addition, all class proportions are summed up to 1 as in Equation 7.5 below:

$$\sum_{l=1}^{L} p_l = 1. \tag{7.5}$$

Linzer (2011) adopted a similar method involving a dummy variable but used a response indicator term instead of the Kronecker Delta term in Equation 7.1. To explain Linzer (2011)'s method, let I be a response indicator matrix the element $I_{ij'r}$ the observed indicator of r^{th} response level for individual i and categorical variable j'. As this stage, if $y_{ij'} = r$, then $I_{ij'r} = 1$. On the contrary, if $y_{ij'} \neq r$, then $I_{ij'r} = 0$. As such, the probability density function for all j' variables and l latent class is expressed as follows:

$$f(Y_i \mid c_l) = f(I_i; \pi_l) = \prod_{j'=1}^{J'} \prod_{r=1}^{R_{j'}} \pi_{j'rl}^{I_{ij'r}}. \tag{7.6}$$

The probability density function across all L latent classes can be expressed as

follows:

$$f(Y_i) = f(I_i \mid \pi, p_l) = \sum_{l=1}^{L} p_l \prod_{j'=1}^{J'} \prod_{r=1}^{R_{j'}} \pi_{j'rl}^{I_{ij'r}}. \tag{7.7}$$

Given estimates of p_l and $\pi_{j'rl}$, $\hat{p}_l$ and $\hat{\pi_{j'rl}}$ respectively, the posterior probability that individual i belongs to latent class l, l_i, conditional on the vector of response indicator for individual i, I_i, is expressed as follows:

$$\hat{P}(l_i \mid I_i) = \frac{\hat{p}_l f(I_i; \hat{\pi}_l)}{\sum_{l'=1}^{L} \hat{p}_{l'} f(I_i; \hat{\pi}_{l'})}, \tag{7.8}$$

where $l_i \in \{1, \ldots, L\}$. The number of parameters in the latent class analysis model is $L[\sum_{j'=1}^{J'}(R_{j'} - 1) + (L-1)]$, which is the total number of response levels and latent classes, subtracted by baseline response levels and latent classes. The log-likelihood function for the latent class analysis model is here expressed by the following:

$$\ln(f(Y)) = \sum_{i=1}^{n} \ln \sum_{l=1}^{L} p_l \prod_{j'=1}^{J'} \prod_{r=1}^{R_{j'}} \pi_{j'rl}^{I_{ij'r}}. \tag{7.9}$$

In order to estimate $\hat{p}_l$ and $\hat{\pi}_{j'rl}$, the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) is implemented. The EM algorithm begins with initial values of $\hat{p}_l$ and $\hat{\pi}_{j'rl}$, which are labelled as $\hat{P}_l^{old}$ and $\hat{\pi}_{j'rl}^{old}$ respectively. The class membership probabilities, $\hat{P}(l_i \mid I_i)$, are calculated in the expectation step by Equation 7.10 with $\hat{P}_l^{old}$ and $\hat{\pi}_{j'rl}^{old}$. In the maximization step, $\hat{P}_l^{old}$ and $\hat{\pi}_{j'rl}^{old}$ are updated by maximizing the log-likelihood function, expressed in Equation 7.11, given the estimated posterior $\hat{P}(l_i \mid I_i)$ obtained in Equation 7.10. The updated estimates are denoted as $\hat{p}_l^{new}$ and $\hat{\pi}_{j'rl}^{new}$ respectively, with the following expressions:

$$\hat{p}_l^{new} = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(l_i \mid I_i), \tag{7.10}$$

and

$$\hat{\pi}_{j'rl}^{new} = \frac{\sum_{i=1}^{n} I_{ij'} \hat{P}(l_i \mid I_i)}{\sum_{i=1}^{n} \hat{P}(l_i \mid I_i)}. \tag{7.11}$$

The estimates and the standard errors of the latent class were derived by Linzer

(2011), who continued building from work of McLachlan and Peel (2000).

A solution for including covariates in the latent class analysis model is to link the class membership probabilities to a regression model. The resultant model is called the latent class regression model. In terms of model construction, Bandeen-roche and Miglioretti (1997) suggested using the proportion of all exponential sum of regression components as a measure of class membership probabilities. More precisely, the component P_l is replaced by a mixing proportion $P_{li} = P_l(X_i;\beta)$, where $X = [X_1,\ldots,X_i,\ldots,X_n]$ is a $n \times R$ covariate matrix, $\beta = [\beta_1,\ldots,\beta_L]^T$ is a $R \times L$ parameter matrix for L latent classes, $R = \sum_{j'=1}^{J'} R_{j'}$ is the number of covariate parameters in the regression model. The mixing proportion is expressed by the following equation:

$$p_{li} = p_l(X_i;\beta) = \frac{e^{X_i\beta_l}}{\sum_{l'=1}^{L} e^{X_i\beta_{l'}}}. \tag{7.12}$$

Combining Equations 7.7 and 7.12, Linzer (2011) derived the probability density function for latent class regression model as follows:

$$f(Y_i) = f(I_i \mid \pi, p_l) = \sum_{l=1}^{L} \frac{e^{X_i\beta_l}}{\sum_{l'=1}^{L} e^{X_i\beta_{l'}}} \prod_{j'=1}^{J'} \prod_{r=1}^{R_{j'}} \pi_{j'rl}^{I_{ij'r}}. \tag{7.13}$$

The log-likelihood function of latent class regression model is expressed in Equation 7.14 as:

$$\ln(f(Y)) = \sum_{i=1}^{n} \ln \sum_{l=1}^{L} \frac{e^{X_i\beta_l}}{\sum_{l'=1}^{L} e^{X_i\beta_{l'}}} \prod_{j'=1}^{J'} \prod_{r=1}^{R_{j'}} \pi_{j'rl}^{I_{ij'r}}. \tag{7.14}$$

Linzer (2011) used the lowest-level latent class, c_1, as a base level and set the consequent vector β_1 to be $\{0,\ldots,0\}$. The entire latent class regression model then measures the log odds of latent class memberships of class 2 to L inclusive

versus class 1. As a result:

$$\ln\left(\frac{p_{2i}}{p_{1i}}\right) = X_i\beta_2,$$
$$\ln\left(\frac{p_{3i}}{p_{1i}}\right) = X_i\beta_3,$$
$$\vdots$$
$$\ln\left(\frac{p_{Li}}{p_{1i}}\right) = X_i\beta_L. \tag{7.15}$$

The general result from Equation 7.15 is equal to Equation 7.12, given the constraint that $\beta_1 = 0$. The posterior (in this chapter, posterior means after considering the likelihood function) class membership probabilities in the latent class regression model are then expressed as follows:

$$\hat{P}(l_i \mid X_i; I_i) = \frac{p_l(X_i;\hat{\beta})f(I_i;\hat{\pi}_l)}{\sum_{l'=1}^{L} p_{l'}(X_i;\hat{\beta})f(I_i;\hat{\pi}_{l'})}. \tag{7.16}$$

The prior probability estimates are applied to each latent class, $c_1, \ldots, c_L$, whereas the posterior probability estimates are applied to every individual $i = 1, \ldots, n$ of the data set.

In many cases, the number of latent classes is more than two (i.e. $L > 2$), hence more than two posterior proportion probabilities for each individual. Since the range of these variables is contained in the $[0, 1]$ interval, each response can be modeled by beta distribution. However, a problem arises since the beta distribution is only capable of modelling two constrained proportions (e.g. p_i and $1 - p_i$). For analyses where three or more constrained proportions should be modelled at once, the Dirichlet distribution is one viable solution. In Section 7.2.3, we discuss the theory of Dirichlet distribution, as well as its application to the latent class analysis.

7.2.3 Dirichlet Distribution

7.2.3.1 Theory

The Dirichlet distribution is an analysis for data sets containing responses that are three or more probabilities that summed up to 1. It serves as a generalisation of Beta distribution for the number of class $L \geq 2$. Maier (2014) provided the following explanation: Suppose for n individuals, there exists L probability responses, each corresponding to l^{th} latent class: $\rho_l, l = 1, \ldots, L$, where $\rho_l \in (0,1)$ and $\sum_{l=1}^{L} \rho_l = 1$, and shape parameters for each class l, denoted as α'_l, let $\rho = \{\rho_1, \ldots, \rho_L\}$ be a set of probability responses for L corresponding latent classes, then the probability density function for Dirichlet distribution, given a set of shape parameters $\alpha' = \{\alpha'_1, \ldots, \alpha'_L\}$, is described by Equation 7.17:

$$\rho \sim D(\alpha'); f(\rho \mid \alpha') = \frac{1}{B(\alpha')} \prod_{l=1}^{L} \rho_l^{\alpha'_l - 1}, \tag{7.17}$$

where the normalising constant $B(\alpha')$ is expressed as follows:

$$B(\alpha') = \prod_{l=1}^{L} \frac{\Gamma(\alpha'_l)}{\Gamma(\sum_{l'=1}^{L} \alpha'_{l'})}, \tag{7.18}$$

and the Gamma function $\Gamma(.)$ is defined as follows:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt. \tag{7.19}$$

Moreover, $\alpha'_l > 0, l = 1, \ldots, L$. By denoting $\alpha'_0 = \sum_{l=1}^{L} \alpha'_l$, with such set of parameters, the mean of Dirichlet distribution for class l is $\mathbf{E}[\rho_l] = \alpha'_l / \alpha'_0$, the variance is $\text{VAR}[\rho_l] = [\alpha'_l(\alpha'_0 - \alpha'_l)]/[(\alpha'_0)^2(\alpha'_0 + 1)]$, and the covariances are $\text{COV}[\rho_l, \rho_{l'}] = (-\alpha'_l \alpha'_{l'})/[(\alpha'_0)^2(\alpha'_0 + 1)]$

Each class l is marginally distributed from beta distribution, $B(\alpha'', \beta')$, with $\alpha'' = \alpha'_l$ and $\beta' = \alpha'_0 - \alpha'_l$. The full log-likelihood of the Dirichlet distribution

with the original parameterisation is defined as:

$$\ln(f(\rho \mid \alpha')) = \log\left[\Gamma\left(\sum_{l=1}^{L}\alpha'_l\right)\right] - \sum_{l=1}^{L}\log[\Gamma(\alpha'_l)] + \sum_{l=1}^{L}(\alpha'_l - 1)[\log(\rho_l)]. \quad (7.20)$$

7.2.3.2 Application

In real data applications, extreme values of probability responses in a Dirichlet distribution model, which are 0 or 1, may exist. If $\rho_l = 0$, then $\log(\rho_l) = -\infty$, on the other hand, $\rho_l = 1$ implies the other proportion probability to be zero, which may lead to $\log(\rho_{l'}) = -\infty, l' = 1,\ldots,L, l' \neq l$. To remedy the situation of extreme values, Smithson and Verkuilen (2006) suggested the following transformation:

$$\rho_l^* = \frac{\rho_l(n-1) + 1/L}{n}. \quad (7.21)$$

The sum of all shape parameters, α'_0, can also be interpreted as a kind of 'precision' parameter. By considering such a precision parameter, as well as the expectation value of the Dirichlet distribution, this interpretation led to the alternative parameterisation defined by Ferrari and Cribari-Neto (2004). In the alternative parameterisation, let $\mu_l = \mathbf{E}[\rho_l]$ be the expectation parameter, and $\gamma = \alpha'_0$ be the precision parameter, then we obtain $\alpha'_l = \mu_l\gamma$ and $\alpha'_0 = \gamma$. As a result of re-parameterisation, the expectation, variance and covariance of Dirichlet distribution between latent class l and $l', l' = 1,\ldots,L, l' \neq l$ are defined as $\mathbf{E}[\rho_l] = \mu_l$, $\mathrm{VAR}[\rho_l] = [\mu_l(1-\mu_l)]/(\gamma+1)$ and $\mathrm{COV}[\rho_l, \rho_{l'}] = -\mu_l\mu_{l'}/(\gamma+1)$, where $\mu_l \in (0,1)$ and $\gamma > 0$. Let $\mu = \{\mu_1,\ldots,\mu_L\}$. The probability density function for the Dirichlet distribution under alternative parameterisation is derived as the following equation:

$$f(\rho \mid \mu, \gamma) = \frac{1}{\mathrm{B}(\mu\gamma)}\prod_{l=1}^{L}\rho_l^{(\mu_l\gamma-1)} \quad (7.22)$$

(Maier, 2014),

where $\mu_l \in (0,1)$ and $\gamma > 0$. Due to the constraints of these parameters, two dif-

ferent links were used for the expectation parameters of all classes and precision parameter respectively: the multinomial logit link for the expectation parameters and log-arithmetic link for precision parameter. These link functions are expressed in the following fashion:

$$\eta_{\mu_l} = \mathbf{X}\beta_l, \tag{7.23}$$

$$\eta_{\gamma} = \mathbf{Z}\gamma. \tag{7.24}$$

Since for each case the sum of all ρ_l is constrained to be 1, the alternative method of parameterisation is to set the lowest class (usually class 1, or generally, class l'', the baseline class) as the reference class and obtain the proportion ratio with respect to the reference class (i.e. $\rho_l/\rho_{l''}$). The mean parameters for the baseline class l'' and class l are denoted as $\mu_{l''}$ and μ_l. The regression coefficients for the baseline class are set to zero (i.e. $\beta_{l''} = \mathbf{0} = (0,\ldots,0)$). As a result of re-parameterisation, the expected values for the baseline class and any class l are modelled as follows:

$$\mu_l = \frac{e^{X\beta_l}}{\sum_{l'=1}^{L} e^{X\beta_{l'}}}, \tag{7.25}$$

$$\mu_{l''} = \frac{e^{X\mathbf{0}}}{\sum_{l'=1}^{L} e^{X\beta_{l'}}} = \frac{1}{\sum_{l'=1}^{L} e^{X\beta_{l'}}}, \tag{7.26}$$

$$\ln[(\rho \mid \alpha')] = \log\Gamma(\gamma) - \sum_{l=1}^{L} \log\Gamma(\alpha'_l) + \sum_{l=1}^{L} (\alpha'_l - 1)(\rho_l). \tag{7.27}$$

Further details about the Hessian matrix can be found in Maier (2014).

7.2.4 Application of Latent Class Analysis

In this research, the `R` program for polychotomous latent class analysis, `poLCA`, was adopted. The `R` function of latent class analysis, `poLCA(.)`, was used for fitting latent class analysis in `R` program (`R` version 3.3.0). Two, three and four latent classes were specified. The `R` program begins at a set of arbitrary class-

conditional probabilities for all drug-trying response variables (unless a random seed is specified). The maximum number of iteration cycles for the EM algorithm was set to 100,000 to guarantee convergence of this Bayesian estimation. In order to avoid local maxima of the log-likelihood, ten sets of estimates were generated, and the set with the lowest log-likelihood was chosen. To generate a consistent result, a random seed number 4321 was used. After modelling, the starting values of the class-conditional probabilities for all drug-trying response variables were sorted, and were then used in executing the `poLCA` command again. Re-running the latent class analysis model led to a change in the ordering of the latent classes by class proportions (e.g. there were three classes in the latent class analysis, *A*, *B* and *C*, class *A* yielded the largest class proportion and class *C* yielded the smallest class proportion. In a latent class analysis model, these three classes were ordered as *B*, *C* and *A*, but after re-running the model, these classes were re-ordered as *A*, *B* and *C*), and hence the ordering of probabilities. However, given that the sufficient iteration cycle and sufficient number of estimating sets were specified, the values of estimates were not affected.

Throughout the analysis, the predicted class memberships for all students, class proportions and class-conditional posterior probabilities for all drug responses, as well as estimates and standard errors for covariates for the latent class regression model, were obtained. The class membership for each case was assigned according to the greatest probability of the posterior probability of each case.

In this research, latent class analysis (among drug-trying response variables only) was performed on the ten imputed data sets resulting from the MICE scheme 2 using `poLCA` package in `R` program. Also, alternative environment used for conducting the latent class analysis was the `Latent Gold` program (Vermunt and Magidson, 2008). The `Latent Gold` program also permits the fitting of a latent class analysis as well as a latent class regression model, to

a data set with missingness. The missing values are by default imputed by a non-parametric bootstrap procedure. An alternative imputation method is via the EM procedure. According to the Latent Gold 4.5 Syntax Manual (Vermunt and Magidson, 2008), the non-parametric bootstrap procedure is preferred because the procedure considers the imputation uncertainty. Whilst the latent class analysis was mainly conducted using `poLCA` package in `R` program, it was also conducted using the `Latent Gold` program via both EM procedure and non-parametric bootstrap procedure as a sensitivity analysis, in order to compare with the results generated from the `R` program.

In order to compare results of the latent class analysis models based on different imputation procedures, both non-parametric bootstrap procedure and EM procedure were considered in the `Latent Gold` program. The results generated by both procedures returned similar values, which implied that the analyses were not sensitive to this change in imputation procedures. The modelling work on latent class analysis and latent class regression model there-after were conducted on the `poLCA` package in `R` program.

Since fitting a latent class regression model with a large number of covariates (for example 30 covariates) was found to be computationally challenging, with three latent classes or more specified, a pre-selection of covariates was needed. The rationale was thus to choose a subset of likely predictors and hence to reduce the computational complexity. In order to pre-select the covariates, firstly latent class analysis was performed on the fifteen drug-trying response variables, with options of latent classes $L = 2, 3, 4$. This was repeated for the $M = 10$ data sets. The optimal number of latent classes was then chosen, based on the lowest BIC (Schwarz, 1978) and adjusted BIC (Sclove, 1987), for each of the ten imputed data sets. The equation for the BIC (denoted as *BIC*) for a tested model as described in Equation 7.28 below is:

$$BIC = -2l + P'' * \ln(n) \tag{7.28}$$

(Schwarz, 1978),

where l represents the log-likehood of the tested model, and P'' represents the number of parameters in the tested model. The equation for the adjusted BIC (denoted as BIC_{adj}) for a tested model as described in Equation 7.29 below is:

$$BIC_{adj} = -2l + P'' * \ln\left(\frac{n+2}{24}\right) \tag{7.29}$$

(Sclove, 1987).

The global optimum was then chosen and is denoted L_{opt}. The next step was to obtain the L_{opt} posterior probabilities of class membership for each student, and these values sum to one. A Dirichlet regression model (alternative parameterisation) with all covariates was then fitted to each of the $M = 10$ imputed data sets. Estimates and standard errors of the covariate coefficients were combined using Rubin's rule. The non-significant covariates in the L_{opt} regression models (according to combined estimates and standard errors) were discarded, one at each step, using backward elimination and a 5 % significance level. All the covariate terms that remained significant in at least one of the L_{opt} models were therefore selected for use in the latent class regression, namely "one-stage latent class regression model". The algorithm for the "one-stage latent class regression model" is given in Algorithm 7.1.

Algorithm 7.1 one-stage latent class regression model

1: Begin with full model (with all covariates), based on ten imputed data sets;
2: **while** Insignificant variables remain in the model **do**
3: **for** $w = 1, \ldots, W$ **do**
4: Fit a two-class, three-class or four-class model with a current set of covariates for data set w;
5: Choose the optimal number of classes based on BIC and "adjusted BIC" for data set w;
6: **end for**
7: for the models with the optimal number of classes, combine all estimates and standard errors by Rubin's Rule, and discard one covariate with the largest p-value (which is greater than 0.05) by Wald's test;
8: repeat Lines 2 to 7 without the covariate discarded at Line 7;
9: **end while**

7.2.5 Results of the Latent Class Analysis Model

Firstly, the values of BIC and adjusted BIC for $L = 2, 3, 4$ latent class analysis models based on the ten imputed data sets are displayed in Table 7.2.1.

Table 7.2.1: Table of BIC and Adjusted BIC for Latent Class Analysis Models Fitted Using `poLCA` Package in the `R` Program (*: lowest value)

	BIC			adjusted BIC		
Data Set	2-class	3-class	4-class	2-class	3-class	4-class
1	16357.12	15996.86*	16069.57	16258.61	15847.51*	15869.37
2	16085.64	15777.18*	15835.27	15987.13	15627.82*	15635.07
3	15961.57	15670.15*	15743.23	15863.06	15520.79*	15543.03
4	16000.5	15731.12*	15793.69	15901.99	15581.76*	15593.49
5	16054.71	15694.21*	15740.53	15956.2	15544.85	15540.33*
6	15898.58	15627.17*	15699.52	15800.07	15477.81*	15499.32
7	15975.38	15704.12*	15763.54	15876.87	15554.76*	15563.34
8	16192.26	15864.19*	15934.65	16093.75	15714.83*	15734.45
9	16206.5	15862.27*	15911.63	16107.98	15712.92	15711.43*
10	16262.77	15910.25*	15983.4	16164.26	15760.9*	15783.2

According to Table 7.2.1, the lowest BIC was observed for all ten imputed data sets for the $L = 3$ class option. In addition, the lowest adjusted BIC was observed in eight out of ten imputed data sets for the $L = 3$ class option. This indicated that the three-class option was, in general, classified the respondents parsimoniously. As such, the latent class analysis proceeded with three classes.

Table 7.2.2 displays the proportions of combined class membership over the $M = 10$ data set, for each of the three classes. Table 7.2.3 lists the student frequencies for the three classes based upon posterior class membership assigned by the greatest class probability for every student, for ten imputed data sets, whereas Table 7.2.4 displays the class-conditional posterior probabilities arising from the latent class analysis, for the models generated using `R` and `Latent Gold` programs, respectively.

Table 7.2.2: Combined Class Membership Proportion Table of Latent Classes for the `R` and `Latent Gold` Programs

Program	Class 1	Class 2	Class 3
`R` (`poLCA`)	0.9265	0.0654	0.0081
`Latent Gold`	0.9352	0.0576	0.0071

From Table 7.2.2, the class membership proportions for both models generated from the `R` and `Latent Gold` programs were similar, with class 1 being the largest group (class proportions of 0.9265 and 0.9352 for the `R` and `Latent Gold` programs respectively), followed by class 2 (class proportions of 0.0654 and 0.0576 for the R and Latent Gold programs respectively). Class 3 was a posteriori as the smallest group. The proportions of class 2 and class 3 for the model generated from the `R` program were larger than those generated from the `Latent Gold` program. On the contrary, the proportions of class 1 for the model generated from the `R` program were slightly smaller than those generated from the `Latent Gold` program.

Table 7.2.3: Predicted Frequency Table for Three-class Latent Class Analysis Model using the R program

	Class 1	Class 2	Class 3
1	6829	410	57
2	6887	357	52
3	6885	352	59
4	6869	372	55
5	6879	360	57
6	6874	366	56
7	6880	354	62
8	6865	371	60
9	6875	363	58
10	6854	383	59

From Table 7.2.3, the three frequencies of students in the corresponding three classes between the ten imputed data sets were also generally similar, albeit the frequencies of students in class 1 and 2 for data set 1 were slightly different, due to the higher percentages of the students trying cannabis and gas in imputed data set 1 when compared with other imputed data sets.

Table 7.2.4: Table of Class-conditional Posterior Probabilities of Latent Class Analysis Models for the R and Latent Gold Programs Without Covariates

	R (poLCA)			Latent Gold		
Variable	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Cannabis	0.0466	0.7434	0.9278	0.0442	0.7185	0.9109
Heroin	0.0004	0.0392	0.4972	0.0003	0.0335	0.3798
Cocaine	0.0005	0.1294	0.7648	0.0004	0.1076	0.6842
Magic Mushrooms	0.0009	0.1692	0.7594	0.0007	0.1534	0.6527
Crack	0.0006	0.0559	0.4971	0.0004	0.0482	0.3900
Methadone	0.0009	0.0546	0.5578	0.0008	0.0472	0.4705
Ecstasy	0.0004	0.1046	0.7862	0.0004	0.0790	0.7516
Amphetamines	0.0000	0.0992	0.6274	0.0000	0.0861	0.5225
LSD	0.0003	0.0441	0.5236	0.0003	0.0352	0.4439
Poppers	0.0021	0.2882	0.7194	0.0016	0.2649	0.6605
Ketamine	0.0007	0.0519	0.4334	0.0005	0.0477	0.3338
Anabolic Steroids	0.0004	0.0510	0.2893	0.0003	0.0468	0.2172
Gas	0.0638	0.3360	0.5619	0.0628	0.3274	0.4879
Other drugs	0.0005	0.0454	0.2656	0.0004	0.0396	0.2298
Tranquillisers	0.0005	0.0310	0.4067	0.0004	0.0273	0.3243

From Table 7.2.4, the class-conditional posterior probabilities for data sets imputed under the R program were all higher than those for data sets imputed

under the `Latent Gold` program, albeit the differences were small. This might due to imputation model differences and different estimation algorithms in these two programs. Table E.1.1 in Appendix E supported this argument by displaying mostly lower frequencies across the ten imputed data sets based on the `R` program than those based on the `Latent Gold` program for cannabis, gas and other drugs, but higher frequencies for other drug-trying response variables.

To assist with interpretation of the drug proportions in Table 7.2.4, Figure 7.2, shows the combined class-conditional posterior probabilities of drug-trying response variables for models generated from the `R` and `Latent Gold` programs.

Figure 7.2: Class-conditional Probability Plot for the Drug-trying Response Variables in Latent Class Analysis Model (Left: R program; right: `Latent Gold` program)

Both Table 7.2.4 and Figure 7.2 compared the class-conditional posterior probabilities of drug-trying response variables across three classes. In Class 1, drug-trying was rare but distinct small probabilities for cannabis and gas were observed. In Class 2, relatively large probabilities for cannabis, poppers and gas, as well as a smaller probability for magic mushrooms were observed. Since the corresponding drugs were soft drugs, Class 2 could be classified as a 'soft drug group'. Class 3 contained relatively large class-conditional probabilities of all drugs and could be classified as 'soft and hard drug group'.

The next stage of analysis was to conduct Dirichlet regression model with backward elimination. 34 covariates that worked with the Dirichlet regression model were included in the initial model. Ten imputed data sets generated from MICE were used. The estimates and covariance matrices of ten Dirichlet regression models were combined by Rubin's rule. In each step of backward elimination, the covariate with the largest p-value that was larger than 0.05 significance level was discarded. This process continued until no insignificant covariates remained in the Dirichlet regression model.

From the resultant dirichlet regression model, the following covariates were chosen: CgStat1, CgPk1, CgGet1, CgGet2, CgPp1, CgBuyF1, CgEstim, CgIn1, AlFreq2, AlBnPub1, AlPar1 (numeric), Al4W1, AlWhy1, DgIn1, DgEstim, Age (numeric), Gender, TruantN, ExclAN1 and SHA. These covariates were found to be significant in any one of the three dirichlet regression models. Combined with nine covariates which were found to be significant in at least three of fifteen logistic regression models: CgStat1, CgWho1, CgBuyF1, CgEstim, AlFreq2, DgEstim, Books1 (numeric), Age (numeric) and TruantN; a total of 22 covariates (7 covariates out of total 29 covariates overlapped per the above two lists) were selected as starting covariates of "one-stage latent class regression model" model (mentioned in Section 7.2.4) at the first step of backward elimination as

follows: CgStat1, CgPk1, CgGet1, CgGet2, CgPp1, CgWho1, CgBuyF1, CgEstim, CgIn1, AlFreq2, AlBnPub1, AlPar1 (numeric), Al4W1, AlWhy1, DgIn1, DgEstim, Books1 (numeric), Age (numeric), Gender, TruantN, ExclAN1 and SHA.

7.2.6 New Methodology: the Algorithm for the Backward Elimination in the Latent Class Regression Model Using Rubin's Rule with Wald's Test

Before discussing the results of the latent class regression model, we present the algorithm for the backward elimination in the latent class regression model, implemented using Rubin's rule with Wald's test.

Steps of Backward Elimination with Wald's Test for Latent Class Regression Model:

a. For w imputed data sets, $w = 1, \ldots, W$, fit a saturated model with the following 22 covariates:
CgStat1, CgPk1, CgGet1, CgGet2, CgPp1, CgWho1, CgBuyF1, CgEstim, CgIn1, AlFreq2, AlBnPub1, AlPar1 (numeric), Al4W1, AlWhy1, DgIn1, DgEstim, Books1 (numeric), Age (numeric), Gender, TruantN, ExclAN1 and SHA

b. i. For w imputed data sets, $w = 1, \ldots, W$, model the 15 drug-trying response variables (also 22 drugs, smoking and drinking indicators) using $L = 2, 3$ latent classes. Fit each model using ten iterations to obtain the model with the maximum global likelihood.

ii. Sort the class proportions in descending order. Sort the starting values of the class-conditional probabilities of all the drug-trying response variable accordingly and re-fit the model with the sorted starting values.

iii. Choose the model with the lowest adjusted BIC amongst the three models.

c. (1) For w imputed data sets, $w = 1, \ldots, W$, choose a variable to discard by Rubin's rule with Wald's Test.

i. remodel the 15 drug-trying response variables (also 22 drugs, smoking and drinking indicators) with all potential covariates.
Fit each model with various latent classes for ten times to obtain the model with the maximum global likelihood.

ii. Sort the class proportions in descending order, sort the starting values of the class-conditional probabilities of all the drug-trying response variables accordingly and re-fit the model with the sorted starting values.

iii. For each covariate that was considered to be discarded, obtain an estimate matrix and a corresponding covariance matrix;

The parameter estimate and the corresponding covariance matrix are then obtained and hence transformed for computational reasons. More specifically, a factor with levels A, B, C and two latent classes comparator levels (i.e. class 2:class 1 and class 3:class 1), (1, 2) for latent class regression model with three classes. The matrices take the form:

$$\begin{pmatrix} \theta_{A1} & \theta_{A2} \\ \theta_{B1} & \theta_{B2} \\ \theta_{C1} & \theta_{C2} \end{pmatrix}, \begin{pmatrix} V_{A1,A1} & V_{A1,B1} & \cdots & V_{A1,C2} \\ V_{B1,A1} & V_{B1,B1} & \cdots & V_{B1,C2} \\ V_{C1,A1} & V_{C1,B1} & \cdots & V_{C1,C2} \\ V_{A2,A1} & V_{A2,B1} & \cdots & V_{A2,C2} \\ V_{B2,A1} & V_{B2,B1} & \cdots & V_{B2,C2} \\ V_{C2,A1} & V_{C2,B1} & \cdots & V_{C2,C2} \end{pmatrix}.$$

(2) The transformation is then made:

$$\begin{pmatrix} \theta_{A1} \\ \theta_{A2} \\ \theta_{B1} \\ \theta_{B2} \\ \theta_{C1} \\ \theta_{C2} \end{pmatrix}, \begin{pmatrix} V_{A1,A1} & V_{A1,A2} & \cdots & V_{A1,C2} \\ V_{A2,A1} & V_{A2,A2} & \cdots & V_{A2,C2} \\ V_{B1,A1} & V_{B1,A2} & \cdots & V_{B1,C2} \\ V_{B2,A1} & V_{B2,A2} & \cdots & V_{B2,C2} \\ V_{C1,A1} & V_{C1,A2} & \cdots & V_{C1,C2} \\ V_{C2,A1} & V_{C2,A2} & \cdots & V_{C2,C2} \end{pmatrix}.$$

(3) Combine estimates and standard error of W data sets, carry out Wald's test for each covariate (i.e. only include estimates and covariates that are not related to intercept and other covariates)

(4) Discard the covariate with the highest p-value (by Wald's test).

d. Repeat the same process (starting from step b) until no insignificant covariates remain in the model.

7.2.7 Results of the Latent Class Regression Model

7.2.7.1 Results of the Initial Latent Class Regression Model with Covariates

Similar to the latent class analysis model without covariates, in this research, the initial latent class regression model with covariates divided students into the same three classes, with the largest class consisting of students who had tried no drugs and those who had tried cannabis or gas only. Similarly, a much smaller class consisted of students who had tried cannabis, poppers, magic mushrooms and gas, and the smallest class who had tried at least three drugs. The largest class 1 could be referred as "no drugs and cannabis or gas users", the second largest class 2 as "soft drug users" and the smallest class 3 as "soft and hard drug users". Figure 7.3 depicts the combined class-conditional posterior probability plot of the division of the three latent classes.

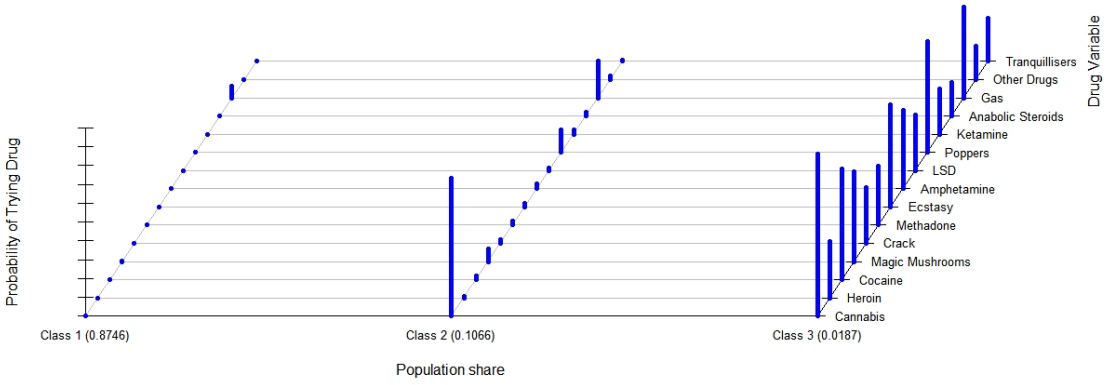

Figure 7.3: Combined Class-conditional Posterior Probability Plot for the Drug-trying Response Variables in the Initial Latent Class Model with Covariates for the `Latent Gold` Program

From Figure 7.3, latent class 1 yielded the smallest class-conditional posterior probability for gas, which was lower than 10 % whereas latent class 2 yielded relatively high class-conditional posterior probabilities for cannabis, magic mushrooms, poppers and gas, with the values between 20 % and 80 %, and latent class 3 yielded high class-conditional posterior probabilities for all drugs, with the values larger than 20 %. In addition, the results generated from the ten data sets were consistent.

Based on the results generated by the Latent Gold program, it was observed that the generated class-conditional prior probabilities and posterior probabilities (after modelling with covariates) were consistent within the ten imputed data sets. On the other hand, when comparing the generated class-conditional prior probabilities with class-conditional posterior probabilities, there were slight discrepancies between them. Such results indicate that both the close associations within drug-trying response variables as well as the associations between

drug-trying response variables and covariates, influenced the class membership proportions of the students in the survey.

7.2.7.2 Results of the Final Latent Class Regression Model with Covariates

In this research, to relate the latent class regression model with covariates, 22 covariates were included in the initial model of the latent class regression model. The backward elimination by Rubin's rule with Wald's test proceeded without altering the number of classes (i.e. between two and three classes, under the BICs and adjusted BICs criteria, the number of classes was chosen to be three). After eleven steps, the final model was produced with 11 covariates. Table 7.2.5 displays the predicted posterior probabilities of the final latent class regression model with covariates and that of the latent class analysis model without covariates, followed by Tables 7.2.6 to 7.2.8, which describe the estimates and standard errors of covariate terms for the final latent class regression models, as well as the BIC table for the initial and the final latent class regression models. The full description of each covariate can be referred to Tables A.2.1 to A.2.3 in Appendix A.

Table 7.2.5: Class Membership Proportion Table of the Final Latent Class Regression Model with Covariates against the Latent Class Analysis Model without Covariates for the R Program

	Class 1	Class 2	Class 3
Without Covariates	0.9265	0.0654	0.0081
With Covariates	0.8746	0.1066	0.0187

Table 7.2.6: Table of Estimates for the Final Latent Class Regression Model (Table 1)

		Class 2 vs 1			Class 3 vs 1		
Variable	Factor	Estimate	SE	P-value	Estimate	SE	P-value
(Intercept)		-7.4432	0.4823	0.0000	-6.7361	1.0695	0.0000
CgStat1	Tried/ Ex-smoker (1)	2.2962	0.2254	0.0000	1.1262	0.6324	0.0749
	Current-light (2)	1.3502	1.0826	0.2123	2.9315	1.1917	0.0139
	Current-moderate (3)	1.7264	0.9147	0.0591	2.9589	1.0631	0.0054
	Current-heavy (4)	2.2727	1.2056	0.0594	5.1313	1.2971	0.0001
CgPk1	Packed (1)	1.5004	1.0461	0.1515	-1.1728	1.1996	0.3282
	Hand-rolled (2)	2.8901	1.0841	0.0077	1.9660	1.2420	0.1134
	Both (3)	2.5328	1.0962	0.0209	-0.1005	1.1802	0.9321
CgGet1	1 (1)						
	> 1(2)						
CgGet2	Shops only (1)						
	1(2)						
	> 1(3)						
CgPp1	Other relatives only (1)						
	Friends only (2)						
	Family members only (3)						
	Mixture (4)						
CgWho1	Smoke, outside (1)						
	Smoke, inside (2)						
CgBuyF1	Few (1)	1.1662	0.2828	0.0000	2.2023	0.4988	0.0000
	Occasional (2)	1.0947	0.3968	0.0058	1.8032	0.5395	0.0008
	Frequent (3)	-0.5600	0.7871	0.4768	0.4676	0.9261	0.6136

Table 7.2.7: Table of Estimates for the Final Latent Class Regression Model (Table 2)

		Class 2 vs 1			Class 3 vs 1		
Variable	Factor	Estimate	SE	P-value	Estimate	SE	P-value
CgEstim	Few (1)						
	Half (2)						
	Most, but not all (3)						
	All (4)						
CgIn1	Passive (1)						
	Interactive (2)						
	Both (3)						
AlFreq2	Ex-drinker (1)	1.6907	0.4669	0.0003	1.2472	0.8643	0.1490
	Few a year (2)	1.5456	0.3220	0.0000	-0.0974	0.8896	0.9128
	Once a month (3)	1.9444	0.3399	0.0000	0.5597	0.8644	0.5173
	Current-Light (4)	2.3141	0.3526	0.0000	0.2691	0.8689	0.7568
	Current-Moderate (5)	2.3058	0.3945	0.0000	0.9293	0.9603	0.3332
	Current-Heavy (6)	2.9914	0.5257	0.0000	2.3663	1.0167	0.0199
AlBnPub	Yes (1)						
AlPar1	(Numeric)	-0.4928	0.1837	0.0073	-0.1126	0.3932	0.7746
Al4W1	Drink, no issue (1)	0.6481	0.2419	0.0074	1.0725	0.7234	0.1382
	Drink, health issue (2)	0.8471	0.3262	0.0094	1.5145	0.8412	0.0718
	Drink, aggressive and other issue (3)	0.5563	0.3597	0.1220	1.6807	0.7357	0.0223
	Drink, Both (4)	0.7401	0.2878	0.0101	2.1587	0.7446	0.0037
AlWhy1	Feel better (1)						
	Socialise (2)						
	Both (3)						

Table 7.2.8: Table of Estimates for the Final Latent Class Regression Model (Table 3)

		Class 2 vs 1			Class 3 vs 1		
Variable	Factor	Estimate	SE	P-value	Estimate	SE	P-value
DgIn1	Passive Media (1)						
	Interactive Media (2)						
	Both (3)						
DgEstim	Only a few (1)	1.0072	0.1984	0.0000	1.0101	0.7330	0.1682
	About half (2)	2.0088	0.2959	0.0000	2.1527	0.7917	0.0065
	Most to all (3)	2.7816	0.4204	0.0000	4.2215	0.8040	0.0000
Books1	(numeric)						
age	(numeric)	0.2563	0.0868	0.0032	-0.2948	0.1707	0.0841
gender	Girl (1)	-1.0919	0.1609	0.0000	-0.7484	0.3576	0.0364
TruantN	Played truant, not in last 12 months (1)	0.7100	0.2767	0.0103	1.3801	0.5527	0.0125
	Once/ twice (2)	0.9355	0.2152	0.0000	0.5265	0.5506	0.3390
	>= 3 times (3)	1.0579	0.3219	0.0010	1.9173	0.4655	0.0000
ExclAN1	Been excluded, not in last 12 months (1)						
	1-2 times (2)						
	>= 3 times (3)						
SHA	North West/Merseyside (1)	0.8932	0.3717	0.0163	-0.2024	0.6636	0.7603
	Yorkshire and the Humber (2)	0.9349	0.3724	0.0121	-0.1857	0.7945	0.8152
	East Midlands (3)	0.2687	0.3392	0.4283	-0.8102	0.6981	0.2458
	West Midlands (4)	0.7297	0.3374	0.0306	-1.0475	0.7820	0.1804
	East of England (5)	0.9143	0.3472	0.0085	-0.5556	0.7742	0.4730
	London (6)	1.4395	0.3848	0.0002	0.2759	0.8258	0.7383
	South East Coast (7)	0.8237	0.3460	0.0173	-0.4404	0.6627	0.5063
	South Central (8)	1.2957	0.3458	0.0002	0.1022	0.6104	0.8671
	South West (9)	0.9848	0.3597	0.0062	-0.0202	0.6655	0.9758

From Table 7.2.5, latent class 1 yielded the dominant proportion in both latent class analysis model (with covariates) and the final latent class regression model, with over 87%, followed by class 2 with 6.54% for latent class analysis model and 10.66% for final latent class regression model respectively. The smallest group was class 3 with a very small proportion of 0.81% and 1.87% for latent class analysis model and latent class regression model respectively.

The class membership proportion of class 1 in the final latent class regression model (with covariates) was lower than that of the latent class analysis model (without covariates). In contrast, the class membership proportions of class 2 and 3 for latent class regression model were greater. This observation indicated that the inclusion of covariates into the final latent class regression model had led to several students being re-allocated from class 1 to mostly class 2 and with some being re-allocated to class 3.

Tables 7.2.6 to 7.2.8 exhibit the relationships between the drug-trying response variables and covariates. A total of eleven covariates remained in the final latent class regression model with covariates: cigarette smoking status (CgStat1), types of smoking (CgPk1), frequency of purchasing cigarettes in shops (CgBuyF1), frequency of drinking (AlFreq2), parents' attitude towards drinking (AlPar1), reasons for drinking (Al4W1), proportion estimate of drug-takers (DgEstim), age of students (between 11 and 15) (Age), gender of students (Gender), frequency of playing truant (TruantN) and Strategic Health Authority (SHA).

Firstly, for the cigarette smoking status (CgStat1) covariate, there was no significant increase in odds ratios of the drug-trying behaviour across level 2 to 4, when comparing latent class 2 to class 1, though the significant odds ratios of level 1 indicated that having a smoking history increased the odds of the students in class 2 trying any soft drug. However, when comparing latent class

3 to class 1, a significant increase in the odds ratios of the drug-trying behaviour across level 2 to 4 was observed. This observation indicated that smoking more heavily led to a higher likelihood of the students in class 3 trying any soft or hard drug.

Secondly, for the types of smoking (CgPk1) covariate, it was found that the students in class 2 who smoked hand-rolled cigarettes were more likely than those in class 1 who did not smoke, by a significant odds ratio of $e^{2.8901} = 17.9951$, to try a soft drug. This coefficient was higher than that representing the students in class 2 who smoked in packed cigarettes, which yielded an odds ratio of $e^{1.5004} = 4.4835$. Moreover, smoking in packed cigarettes appeared having a mitigating effect on the likelihood of the students in class 2 trying a soft drug, since the third level of the CgPk1 variable, representing the students who smoked in both packed and hand-rolled cigarettes, yielded an odds ratio of $e^{2.5328} = 12.5887$, which laid between odds ratio of $e^{2.8901} = 17.9951$ and $e^{1.5004} = 4.4835$. Types of smoking variable did not have any significant effect on the students in class 3 to try any hard or soft drug, at 5 % significance level.

For the frequency of purchasing cigarettes in shops (CgBuyF1) covariate, the students in both class 2 and class 3 who purchased cigarettes a few times (Level 1) or occasionally (level 2) were more likely to try any soft or hard drug than those who did not, as shown by the significant odds ratios of $e^{1.0947} = 2.9883$ and $e^{2.2023} = 9.0458$ respectively. Moreover, the students in class 3 had stronger effects than those in class 2, indicating that the students who purchased more often were more prone to trying more soft or hard drugs. However, as purchasing cigarettes were more often (i.e. beyond level 1 and level 2), the students in both class 2 and class 3 were less likely to try any soft or hard drug than those who purchased cigarettes a few times or occasionally.

For the frequency of drinking (AlFreq2) covariate, estimates from the students in class 2 showed a gradually increasing trend in the likelihood of trying any soft drug, from significant odds ratios of $e^{1.6907} = 5.4233$ to $e^{2.9914} = 19.9135$, when drinking behaviour became more serious. For the students in class 3, no apparent trend was observed from level 1 to level 5, only the drinkers who drank more than once per week and three to seven days in the previous week (heavy drinkers, identified by class 6) were more likely than non-drinkers to try any soft or hard drug, by a significant odds ratio of $e^{2.3663} = 10.6579$.

For the parents' attitude towards drinking (AlPar1) covariate, those students in class 2 who received more encouragement from parents to drink were less likely to try a soft drug. The students in class 2 appeared less likely than those students in class 3 to try a soft drug, with the odds ratio of $e^{-0.4928} = 0.6109$ for the students in class 2, compared to the log odds ratio of $e^{-0.1126} = 0.8935$ for the students in class 3.

For issues associated with the drinking (Al4W1) covariate, the students in both class 2 and class 3 who had drunk in the last four weeks were more likely to try any soft or hard drug, by odds ratios ranging from $e^{0.5563} = 1.7442$ to $e^{2.1587} = 8.6599$. There was no apparent difference in the reason factors for the students in class 2 trying any soft drug, as the estimates were largely similar, between odds ratios of $e^{0.5563} = 1.7442$ and $e^{0.8471} = 2.3329$. Among the students in class 3 who had drunk, those who had no issues related to drinking appeared the least likely to try any soft or hard drug at an odds ratio of $e^{1.0725} = 2.9227$, followed by those who had health issues, with an odds ratio of $e^{1.5145} = 4.5471$. In addition, the students who had become aggressive or had experienced other issues were slightly more likely to try any soft or hard drug, at a significant odds ratio of $e^{1.6807} = 5.3693$, and those who had both issues were the most likely to try any soft or hard drug, at a significant odds ratio of $e^{2.1587} = 8.6599$.

On the other hand, the estimated proportion of known peers who took drugs (DgEstim) covariate, the students in both class 2 and class 3 who reported higher proportions were more likely to try any soft or hard drug. Higher estimates for the class 3 students than those class 2 students indicated that the class 3 students were more likely to try more soft drugs if they knew people who took drugs.

From the Age and Gender covariates, older students in class 2 were more likely to try any drug, at a significant odds ratio of $e^{0.2563} = 1.2921$. Boys were more likely than girls in both class 2 and class 3 to try any soft or hard drug as well. For the frequency of playing truant (TruantN) covariate, generally speaking, students in both class 2 and class 3 who played truant more often were more likely to try any soft or hard drug. The students in class 3 who played truant for at least three times were the most likely to try a soft or hard drug, with a significant odds ratio of $e^{1.9173} = 6.8026$.

Finally, the odds ratio estimates of the Strategic Health Authority (SHA) covariate revealed that London students were most likely to try any soft drug (a significant odds ratio of $e^{1.4395} = 4.2186$ for 'Class 2 vs 1' coefficient), when compared to the students from North East. At the same time, London students appeared most likely to try any soft or hard drug (odds ratio of $e^{0.2759} = 1.3177$ for 'Class 3 vs 1' coefficient). East Midlands students appeared more likely to try any soft drug (odds ratio of $e^{0.2687} = 1.3083$ for 'Class 2 vs 1' coefficient), when compared to students from North East. West Midlands students appeared the least likely to try any soft or hard drug (odds ratio = $e^{-1.0475} = 0.3508$ for 'Class 3 vs 1' coefficient). Across the entire covariate, the class 2 students from North East region were less likely to try any soft drug than the students from the other regions. On the other hand, students from London and South Central were

more likely to try any soft drug than the students from North East.

The main covariate results of the final latent class regression model with covariates could be summarised as follows:

(1) for the smoking covariates: the students who smoked more heavily and more often were more likely to try soft or hard drugs; smoking hand-rolled cigarettes played a more important role than packed cigarettes in influencing the students to try soft drugs and the students purchasing cigarettes a few times were more likely to try soft or hard drugs, but the likelihood faded with an increase in the frequency of cigarette purchase.
(2) For the drinking covariates: the students who drank more heavily were more likely to be subject to drug-trying behaviour; more encouragement from parents to drink led to lower likelihood for the students to try soft drugs. Also, students having drunk in the last four weeks and were involved in both aggressive and health issues and other issues were more likely to try soft or hard drugs.
(3) For the drug-related socio-demographic covariates: the students who knew a larger proportion of people taking drugs were influenced by these people, hence they were more likely to try soft or hard drugs; older boys were more likely than younger girls to try drugs; students who had played truant more often were more likely to try many drugs; and finally, the students from London were the most likely to try drugs.

Apart from discussing the estimates and standard errors of the final latent class regression model, we also discussed in this section the values of BIC, adjusted BIC and AIC of the final latent class regression model, with either two classes or three classes, in order to confirm the three-class option was the best option. We did not fit the latent class regression models with four classes or more, due to excessive computational complexity, which caused the process nearly not to

progress at all. The BIC and adjusted BIC of the final latent class regression model, is presented in Table 7.2.9 below.

Table 7.2.9: Table of BIC and Adjusted BIC of the Final Latent Class Regression Model Across Ten Imputed Data Sets

Data Set	BIC		Adjusted BIC	
	2-class	3-class	2-class	3-class
1	14710.49	13964.04	14268.05449	13175.35209
2	14464.88	13756.91	14022.45049	12968.22409
3	14280.02	13612.68	13837.58649	12823.99009
4	14274.7	13656.75	13832.26649	12868.06009
5	14413.56	13611.97	13971.13049	12823.28609
6	14212.92	13548.4	13770.48049	12759.70809
7	14302.05	13634.92	13859.61449	12846.23209
8	14550.05	13813.42	14107.61849	13024.73609
9	14575	13795.36	14132.56649	13006.67409
10	14151.19	13846.63	14184.51249	13057.94209

Table 7.2.9 showed that final latent class regression model with the three latent classes that had lower BIC and adjusted BIC than the models with the two latent classes. Actually, throughout all the steps in the backward elimination, the latent class regression models with three latent classes were always chosen instead of the models with two latent classes. Latent Class regression models with four or more latent classes, in this case, were computationally challenging to fit.

The combined class-conditional posterior probabilities for the drug-trying response variables in the final latent class regression model, based on ten MICE-imputed data sets, were generated, with several of them being plotted and displayed in Figure 7.4 below.

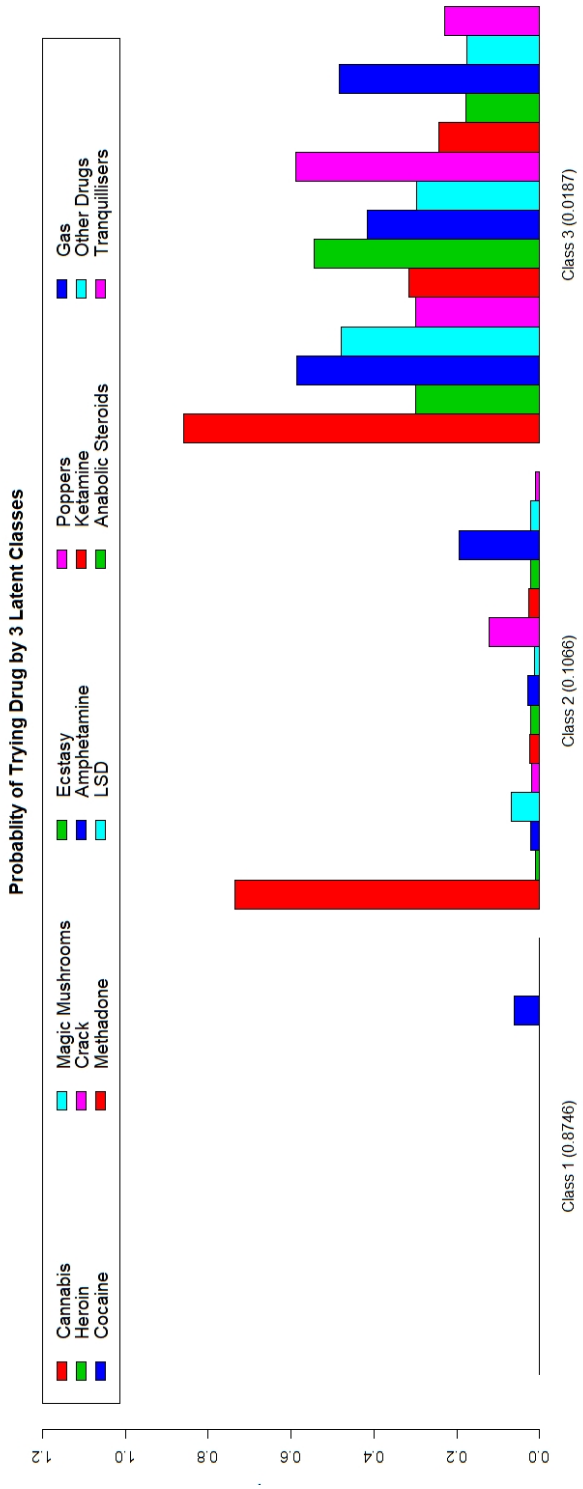

Figure 7.4: Combined Class-conditional Posterior Probability Barplots for the Drug-trying Response Variables in the Final Latent Class Regression Model for the R Program

From Figure 7.4, the class-conditional posterior probabilities in the final latent class regression model with covariates were found to be similar with those in the latent class analysis model without covariates and the initial latent class regression model with covariates, as seen from Figures 7.2 and 7.3 respectively, meaning that the grouping of the three classes, namely (1) no drugs and cannabis or gas users; (2) soft drug users and (3) soft and hard drug users were similar, regardless of the inclusion of covariates.

Connecting Figure 7.4 with Table 7.2.5, class 1 represented the class that comprised of more than 87 % of the students, most of them had not tried any drug but had tried cannabis or gas. Class 2 comprised of about 10 % of the students, most of them had tried cannabis before (with a class conditional posterior probability of more than 75 %) and some of them had tried other soft drugs such as poppers, magic mushrooms and gas. Those in this class did not tend to use hard drugs such as heroin, methadone or crack. Class 3 represented about 1 % of the students who were soft drug or hard drug users.

7.2.8 Discussion and Limitation

Latent class analysis and latent class regression model provided an alternative perspective for investigating drug-trying behaviours. These models were useful as we were able to identify a number of sub-groups and factor classes predicting membership. Specifically, the latent class regression model allowed for classification and regression of latent behaviour on to potential predictors.

In this section, we discuss the possible explanation for the significant covariates resulting from the final latent class regression model. We also contrast the predictors of drug behaviour with those found in the Year 2010 Survey Report, summarised in Section 1.3.

Firstly, we considered the smoking covariates. One possible reason that may explain the positive association between the students who smoked more heavily and more often and the tendency for the students to try soft or hard drugs is that smoking more heavily led to more exposure to drugs. Table 5.2.3 in respect of the univariate logistic regression in Section 5.2 suggested that students who knew smoking peers were more likely to try cannabis. From Tables 5.2.5 and 5.2.6, trying cannabis was found to be associated with trying many other drugs such as cocaine, magic mushrooms and poppers. Another possible reason is that cigarettes contained an addictive called nicotine which has been proven to be a preceding substance of cocaine. With the findings from Tables 5.2.5 and 5.2.6 that trying cocaine was associated with trying many other drugs such as heroin, it is likely that the students will become addicted to hard drugs as long as they smoke regularly. The finding of the positive associations between the cigarette smoking status in general and the students' drug-trying behaviour confirmed the finding in Fuller et al. (2011) Report, as well as the univariate logistic regression findings in Chapter 5. In addition, the odds ratio for the class 2 students who were ex-smokers was higher than that for the class 3 who were ex-smokers, but the odds ratios for the class 2 who were current smokers were lower than that for the class 3 who were current smokers. These findings suggest two further points: (1) regardless of when the students start to smoke, they are more likely than non-smokers to try soft drugs and (2) current smokers are more inclined to try soft or hard drugs than ex-smokers.

Unlike packed cigarettes, for hand-rolled cigarettes, smokers need to fill in a cigarette paper with tobacco and roll it by hand. In this way, smokers who are also drug users can gain more control in the size, the density and the type of tobacco as well as drug powder they are going to smoke. This may explain the students who smoke hand-rolled cigarettes are more likely to try soft drugs.

Also, heavy smokers can pack more tobacco in a single piece of paper to form a cigarette. These heavy smokers of hand-rolled cigarettes are positively associated with the tendency to try soft and hard drugs as discussed in the previous paragraph. Hence, to consider how to mitigate drug abuse, the focus should be more on hand-rolled cigarette smokers than packed cigarette smokers and non-smokers.

The finding that the students who purchased cigarettes quite often were more likely to try soft or hard drugs may be explained by the positive association between the students who smoked more often and the tendency for them to try soft or hard drugs as discussed in the previous paragraph. However, for the finding that if the frequency of purchasing cigarettes was more often then the students were less likely to try soft and hard drugs, the plausible reason may be these students are more obsessed with smoking rather than taking drugs. The similar trend was observed in the univariate logistic regression in Chapter 5 for CgBuyF1 covariate with cannabis or amphetamines as response variables.

Secondly, we considered the findings for the drinking covariates. The finding for the frequency of drinking alcohol indicated that if the class 2 students drank more heavily, they would be more likely to try soft drugs. This finding resembles the result of the logistic regression in (Fuller et al., 2011) report that the frequency of drinking was associated with trying drugs in the previous week. It also agreed to that found in the univariate logistic regression in Section 5.2, which suggested that the higher frequency of drinking contributed to more students trying cannabis, magic mushrooms and poppers. This trend was not clearly seen for the class 3 students, perhaps due to sparse data available in class 3. However, the result that the class 3 students who drank at least once a week and three to seven times in the previous week were more likely than other students to try soft or hard drugs matches with the generally positive association

between the frequency of drinking alcohol and drug-trying behaviour.

The result about parents' attitude towards drinking may reflect a situation that if the parents can provide more support to the students then they may be less likely to try drugs. One possible reason is that the students would be diverted to consume alcohol instead of trying drugs. This result supported with the finding in the univariate logistic regression in Section 5.2 that parents' attitude to drinking was negatively associated with trying cannabis. The degree of drop in likelihood for the class 2 students was larger than that for the class 3 students. This finding can be explained by the understanding that the students, who had tried many types of drugs, are generally more unwilling to abandon their desire to try drugs than those class 2 students who had tried fewer types of drugs.

Then we considered the results of the issues that the students came across when they drank in the last four weeks. Here, the similar odds ratio estimates for the class 2 and class 3 students indicated that, regardless of any issues which the drinkers had, those who had drunk four weeks prior to the survey were more likely to try any soft or hard drug. This finding resembles the finding from the frequency of drinking that generally there was a positive association between the frequency of drinking alcohol and drug-trying behaviour. Results which showed that the class 3 students who had at least one issue were significantly likely to try soft or hard drugs may need more concern.

Generally speaking, alcohol control for adolescents in England is essential if the government officials wish to reduce drug-taking activities. Relying on parents is not sufficiently effective in reducing drug-taking activities.

The positive association between the estimated proportion of peers who took drugs (DgEstim1) and the drug-trying behaviour, for the class 2 and the class

3 students, confirms the findings in the univariate logistic regression in Section 5.2 that cannabis and magic mushrooms were positively associated with the DgEstim1 variable. A plausible reason is the peer influence on drug-trying behaviour, with the students in class 3 who had tried many soft or hard drugs are more influenced by their drug-taking peers.

For the Age covariate, the positive odds ratio for the class 2 students matches with the result found in the Year 2010 Survey Report that has been mentioned in Section 1.3.3. It also matches with the findings in the univariate logistic regression in Section 5.2 that cannabis and magic mushrooms were positively associated with the Age variable. This result indicates that the older students are more likely to try soft drugs. On the other hand, despite insignificant odds ratio for the class 3 students, it showed the negative odds ratio which suggests that the older students are less likely to try many soft or hard drugs. This result also matches with the finding in the univariate logistic regression that Age was negatively correlated with the LSD and gas. One possible explanation is that the younger students may have tried drugs had tried several of them. However, as they grow older, they are more inclined to abandon LSD and gas. Instead, they shift their attention to cannabis and/or magic mushrooms and concentrate on fewer types of drugs.

The finding that girls were less likely than boys to try drugs is consistent with Fuller's results (Fuller et al. (2011), Fuller and Hawkins (2014)). However, the smaller negative odds ratio for the class 3 students indicated that gender is a less important factor for the drug triers who have tried many drugs.

In addition, the odds ratio results of the positive association between frequency of truancy (TruantN) and drug-trying behaviour resemble the result in Fuller et al. (2011) as mentioned in Section 1.3.3 that the frequency of truant was pos-

itively associated with drug-trying behaviour. Despite the insignificant result in level 2 in respect of the odds ratios for the class 3 students, all the odds ratio estimates supported this explanation. The insignificant result for the class 3 students may due to sparse data available in class 3, which results in a rather large standard error, rendering the estimate to be insignificant. The higher odds ratios for the class 3 students compared with those for the class 2 students indicated that playing truant more seriously can lead to students trying more soft or hard drugs.

Finally, the inclusion of the Strategic Health Authority (SHA) in the final latent class regression model produced a new finding that results from both the univariate logistic regression in Section 5.2 of Chapter 5 and the logistic regression that the researchers in Fuller et al. (2011) report, as described in Section 1.3.3 of Chapter 1, agree with each other. In this part, the higher likelihood for the students from London to try soft or hard drugs can be explained by dense population and a greater degree of urbanization in the London area, packing drug-related activities as well as closer purchasing points within London. In contrast, the relative lower likelihood for the students from East Midlands to try soft drugs can be explained by the relatively less dense population and a large area of countryside that provides more space for the students to pursue outdoor activities.

When comparing all the standard errors in Tables 7.2.6 and 7.2.7, the standard errors for class 3 students were much larger than those for class 2. This can be explained by the much smaller population in class 3 than class 2 that more significant estimates were found for class 2 responses than class 3 responses.

However, having considered many measures when running latent class regression models, the following limitations were identified:

(1) If the positive response rate was too low, the result might fluctuate wildly, and the estimates might be unstable.

(2) The latent class analysis was computationally intensive. It was computationally impossible to include all the smoking, drinking and drug-related socio-demographic covariates in the latent class analysis with backward elimination. We needed to pre-select those covariates for the latent class analysis with backward elimination.

The latent class analysis provides a plausible classification of the students based on their drug-trying patterns for this research. In order to provide another perspective on classification of the students, we conducted K-means clustering, which is discussed in Section 7.3 below.

7.3 K-means Clustering

K-means clustering is an alternative algorithm of latent class analysis in stratifying the students based on the pattern, without connection to other covariates. In this research, we compare the results of our latent class analysis with those of K-means clustering, in order to check the validity of classification of the students. We employ K-means clustering to partition all 7,296 students, in order to group the students who are close to some others to form clusters. In the next section (Section 7.3.1), we introduce K-means clustering and explain how it is carried out in this research. Results of K-means clustering are discussed in Section 7.3.4.

7.3.1 Introduction

Clements (1954) suggested the idea of data clustering when dealing with an anthropological data set. Over more than fifty years, data clustering is ubiq-

uitously applied to a wide range of disciplines, for example, statistics, social science, biology and medical research, for its little requirement of assumption.

In general, K-means clustering aims to allocate respondents into clusters by minimising the total square error between each respondent and the cluster mean point for each corresponding cluster (Jain et al., 1999).

Data clustering is an unsupervised classification (or intrinsic classification) of data pattern, since no category labels denoting *a priori* partition of respondents are employed (Jain et al. (1999); Jain and Dubes (1988)). Clustering algorithms can be generally classified into hierarchical clustering and partitional clustering respectively (Jain, 2010). Hierarchical clustering involves clustering responses into a nested sequence of groups, whereas partitional clustering involves splitting responses into separate clusters (Jain and Dubes, 1988).

7.3.2 Theory

K-means clustering is a process of partitioning the N cases into the K clusters in an efficient way, in the sense of within-cluster variance among the J response variables (MacQueen, 1967). Four k-means algorithm options are included in a K-means clustering function, `kmeans(.)`, in the `R` program, (Hartigan and Wong (1979); MacQueen (1967); Lloyd (1982); Forgy (1965)). Lloyd's method can be regarded as Voronoi iteration, generating Voronoi tessellation. However, Telgarsky and Vattani (2010) suggested that Hartigan and Wong's method provided better performance on the synthetic data in the paper than Lloyd's method. The `R` manual regarding the `kmeans` function suggested that Hartigan and Wong's method is better than the other three methods mentioned.

Hartigan and Wong (1979)'s method is based on the K-means clustering al-

gorithm described by Hartigan (1975). Suppose there are n cases, and there are J variables. Euclidean distances are adopted for measuring distances between a data point and its corresponding centroid. Suppose each i^{th} case of j^{th} variable ($i = 1,\ldots,n$, $j = 1,\ldots,J$) yields a value $A(i,j)$, the partition $P(n,K)$ is composed of each cluster $1,\ldots,K$ for each case $1,\ldots,n$ in each cluster $k = 1,\ldots,K$, the mean of the j^{th} variable over all cases in k^{th} cluster is denoted by $B(k,j)$, and the number of cases in cluster k is $n(k)$, the distance between the i^{th} case and k^{th} cluster is expressed as follows:

$$D(i,k) = (\sum_{1\leq j\leq J} [A(i,j) - B(k,j)]^2)^{1/2}. \tag{7.30}$$

The partitioning error term, which measures the sum of distances between every point i and its corresponding mean point, is expressed as follows:

$$e[P(n,K)] = \sum_{1\leq i\leq n} D[i,k(i)]^2, \tag{7.31}$$

where $k(i)$ is the cluster including the i^{th} case. To minimise the partitioning error by general searching procedure, a portion of respondents are reallocated from one cluster to another. The procedure ends when no such movement reduces the error, where the lowest partition error is achieved. The procedure for Hartigan (1975)'s algorithm is listed as follows:

Step 1: Assume initial clusters $1,\ldots,K$. Then compute $B(k,j)$, for $1 \leq j \leq J$ and $1 \leq k \leq K$ and the initial partitioning error as:

$$e[P(n,K)] = \sum_{1\leq i\leq n} D[i,k(i)]^2, \tag{7.32}$$

where $D[i,k(i)]$ denotes the Euclidean distance between i and the cluster mean of the cluster containing i.

Step 2: For the first case, compute, for each cluster k,

$$\frac{n(k)D[1,k]^2}{n(k)+1} - \frac{n[k(1)]D[1,k(1)]^2}{n[k(1)]-1}.$$

This term refers to the increase in error in transferring the first case from cluster $k(1)$ to cluster k. Whenever the minimum of this quantity over all $k \neq k(1)$ is negative, transfer the first case from cluster $k(1)$ to the minimal cluster k, adjust the cluster means of $k(1)$ and the minimal k, then add the increase in error (the error term is negative) to $e[P(n,K)]$.

Step 3: Repeat Step 2 for i^{th} case, for $2 \leq i \leq n$.

Step 4: This procedure ends if no movement of any case from one cluster to another; otherwise, return to Step 2.

7.3.3 Application

In this research, we applied K-means clustering to the 15 drug-trying response variables using all 7,296 students. The clustering was applied to each of the ten imputed data sets, resulting in ten K-means clustering model outputs. The K-means clustering was implemented in the `R` program using the `kmeans` function. It aimed to allocate all the students into K clusters, such that the total sum of squares of Euclidean distances, from each point to its corresponding assigned cluster, was minimised. Since the `kmeans` function in `R` program chose the clustering model generated from G random iterations started with the least sum of squares, it was the best to specify a multiple number of random iterations started by specifying the option in the `kmeans` function, with the syntax "`nstart=G>2`". In this research, G was set to be 1000 to ensure convergence and consistent results, even when adopting different random seeds. Also, a random

seed number 4321 was used for generating results. One to eight-cluster models ($K = 1,2,3,4,5,6,7,8$) were investigated. In order to select the optimal number of cluster, the "Elbow" method was adopted, in which a significant turning point on the total sum of squares graph was identified and the corresponding optimal number of cluster was identified (Ketchen and Shook, 1996).

7.3.4 Results

When adopting the "Elbow" method, a convex curve is generated in Figure 7.5 for K-means clustering models with one to eight clusters, which revealed the most balanced point for the number of clusters.

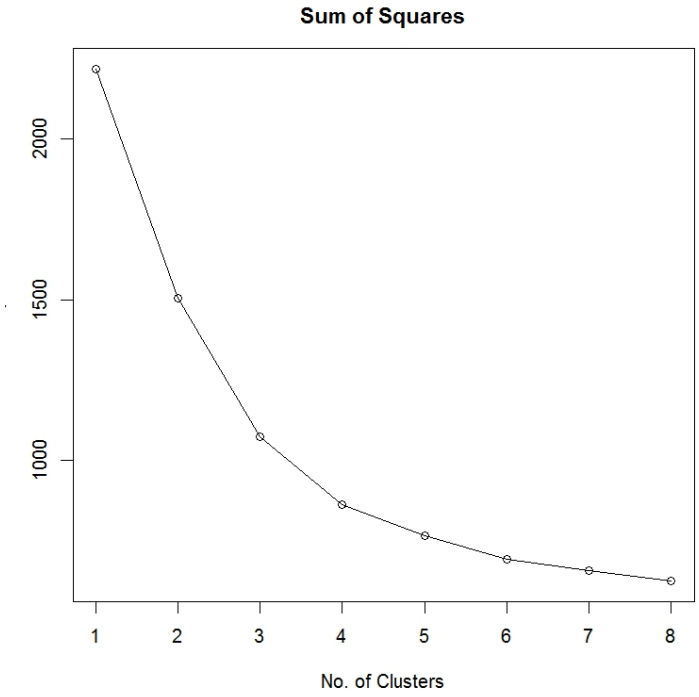

Figure 7.5: Sum of Squares Graphs for K-means Clustering Models with Different Number of Clusters

From Figure 7.5, the graph yielded a significant turning point at four clusters, where adding more clusters to K-means clustering model might lead to diminishing returns. Therefore, the four-cluster K-means clustering model ($K = 4$)

was chosen.

The frequency table and the percentage table for the four-cluster K-means clustering model are displayed in Table 7.3.1 and 7.3.2 respectively:

Table 7.3.1: Frequency Table for Four-Cluster K-means Clustering Model Across Ten Imputed Data Sets

	K-means Cluster			
Data Set	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	6138	462	607	89
2	6141	459	597	99
3	6143	460	591	102
4	6141	465	590	100
5	6145	453	598	100
6	6147	451	595	103
7	6145	457	588	106
8	6144	459	586	107
9	6141	457	597	101
10	6137	459	599	101

Table 7.3.2: Percentage Table (%) for Four-Cluster K-means Clustering Model Across Ten Imputed Data Sets

	K-means Cluster			
Data Set	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	84.13	6.33	8.32	1.22
2	84.17	6.29	8.18	1.36
3	84.20	6.30	8.10	1.40
4	84.17	6.37	8.09	1.37
5	84.22	6.21	8.20	1.37
6	84.25	6.18	8.16	1.41
7	84.22	6.26	8.06	1.45
8	84.21	6.29	8.03	1.47
9	84.17	6.26	8.18	1.38
10	84.11	6.29	8.21	1.38

From Table 7.3.1, the first cluster was the dominant group, followed by the third cluster, then the second cluster and finally the fourth cluster. According to Table 7.3.2, over 84% of the students were in the first group, 6% of the students were in the second group, 8% in the third group and more than 1.2% in the fourth group. The bar plots for cluster-conditional probabilities is displayed in Figure

7.6.

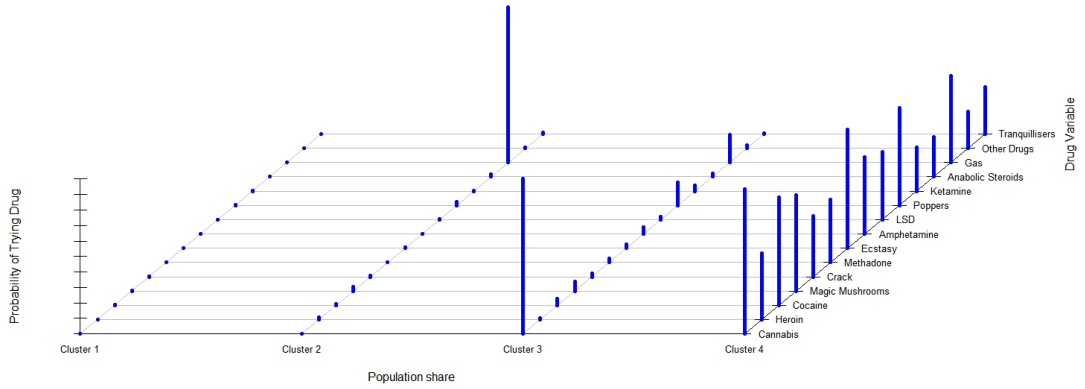

Figure 7.6: Cluster-conditional Probability Bar Plot for Drug-trying Response Variables in K-means Clustering Model for Data Set 1

From the bar plots in Figure 7.6, four clusters with different cluster-conditional probability patterns were observed. Cluster 1 consisted of a majority of the students (about 84 %) who had not tried any drug; Cluster 2 consisted of about 6 % of the students who mostly had tried gas only. At the same time, Cluster 3 consisted of approximately 8 % of the students who had tried cannabis, poppers and gas. The cluster-conditional probability of trying Cannabis was much higher than those cluster-conditional probabilities for poppers and gas. Cluster 4 consisted of barely more than 1 % of the students who had tried many drugs. Another observation from Figure 7.6 was the relatively high cluster-conditional probabilities of trying: (1) cannabis, of which the usage was concentrated in clusters 3 and 4, and (2) gas, of which the usage was concentrated in clusters 2 and 4.

The four clusters modelled by K-means could be interpreted as: (1) the largest cluster that generally consisted of the students who had not tried any drug (cluster 1); (2) the third largest cluster that generally consisted of the students who

mostly had tried gas only (cluster 2); (3) the second largest cluster that generally consisted of the students who had tried soft drugs (i.e. cannabis, poppers and gas) (cluster 3) and (4) the smallest cluster that generally consisted of the students who had tried soft and hard drugs (cluster 4). In general, the clusters were assigned according to the types of drugs the students had tried.

7.3.5 Limitation

According to Santini (2016), K-means clustering is easy to implement and when compared to hierarchical clustering, it requires less computational time. However, in running K-means clustering model, the following limitations were identified:

(1) K-means clustering only clustered cases according to the Euclidean distances between data points of students and their mean point. For data sets with only binomial response variables, regardless of the drug-trying pattern of a student, all cases would be treated the same, given their Euclidean distances are the same.

7.4 Comparison of Latent Class Analysis and K-means Clustering and Discussion

A three-class latent class analysis model without covariates and a four-cluster K-means clustering model were compared in two aspects: (1) group determining method and (2) group assignment.

7.4.1 Group Determining Method

In the latent class analysis, the EM algorithm was adopted, which maximised the log-likelihood of the latent class analysis model and the latent class regression

model. The probability mass function of the latent class analysis model was also taken into account, and the drug-trying pattern of each student was therefore assessed. In other words, different drug-trying patterns of the students led to different probability mass functions, thus leading to different clustering results. On the other hand, K-means clustering employed Euclidean distance as one of the clustering criteria. The main aim of K-means clustering was to minimise the total Euclidean distances between the cluster centroids and case points regardless of the students' drug-trying patterns. In other words, the students of same Euclidean distance would be treated in the same case and would be allocated to the same cluster. To conclude, the latent class analysis took into account the drug-trying patterns of the students in determining which cluster the students were allocated to, whereas K-means clustering did not, but rather allocated the students according to their Euclidean distances.

On the other hand, in the latent class analysis, the optimal number of latent classes was easily identified by the lowest BIC or adjusted BIC value. However, in determining the optimal number of clusters for K-means clustering, we could only resort to a more judgmental method, such as "Elbow Method", where a point of number of clusters should be chosen if a diminishing return was detected beyond such point.

7.4.2 Group Assignment

After running the latent class analysis model, the latent class regression model and the K-means clustering model, similarities in class versus cluster behaviours were examined. When comparing Figure 7.2 with Figure 7.6, it was observed that Class 1 of the latent class analysis model and latent class regression model was basically split into clusters 1 and 2 of the K-means clustering model, where the majority of class 1 members were the students who had not tried any drug

and the much smaller cluster 2 comprised of 6.18 - 6.33% of the students who mostly had tried gas. There were also several students from class 1 of latent class analysis model who were slotted in either cluster 3 or 4 of K-means clustering model. Class 2 of the latent class models was corresponded to cluster 3 of K-means clustering model, due to the similar class/cluster-conditional probabilities of trying any soft drug, where the class/cluster-conditional probabilities for cannabis, poppers and gas were relatively high. Similarly, class 3 of the latent class models was corresponded to cluster 4 of K-means clustering model, due to the similar probability profile of high class/cluster-conditional probabilities of trying soft or hard drugs.

In general, comparing Table 7.3.1 with Table 7.2.3, the total frequency of students in clusters 1 and 2 in K-means clustering model was slightly smaller than the frequency of class 1 in the latent class analysis models. At the same time, the frequency of students in cluster 3 in K-means clustering model was slightly greater than the frequency of class 2 in the latent class analysis models, whereas the frequency of students in cluster 4 in the former was smaller than that of class 3 in the latter. The aforesaid slight frequency discrepancies can be explained by the reason that a small proportion of the students in classes 1 and 3 in the latent class analysis models may be allocated to cluster 3 in K-means clustering model, due to shorter Euclidean distance to the centroid of cluster 3. This reflects the discrepancy of individual predicted class membership from maximum likelihood estimates to cluster allocation based on the shortest total Euclidean distance. Despite the slight frequency discrepancies issue as mentioned above, basically, the K-means clustering model supported the validity of classification of the students in the latent class analysis models.

Generally speaking, though K-means clustering provides a more widespread allocation of students, latent class analysis was considered as a more robust

allocation method in modelling drug-trying behaviour, due to a greater consideration of data pattern in deciding the allocation of groups.

7.5 Summary

In this chapter, the latent class analysis classified the students in accordance with their drug-trying patterns into three classes: (1) class 1: the students tended to not have tried any drug except for cannabis or gas; (2) class 2: the students were more likely to have tried soft drugs (e.g. cannabis, magic mushrooms, poppers or gas) only (3) and class 3: the students had increased tendency to have tried both soft drugs and hard drugs.

The selected 22 covariates were included into the latent class regression model which attempted to explain the relationship between the students of different classes with respect to their drug-trying patterns and the smoking, drinking and drug-related socio-demographic factors via a latent variable. The latent class regression model was conducted by backward elimination by Rubin's rule with Wald's test.

The final latent class regression model revealed that nine covariates were positively associated with the drug-trying behaviour of the students. These nine covariates were: frequency of smoking (CgStat1), type of smoking (CgPk1), frequency of purchasing cigarette (CgBuyF1), frequency of drinking (AlFreq2), reason of drinking (Al4W1), proportion estimate of drug-takers (DgEstim), frequency of truant (TruantN), age (Age) and Strategic Health Authority (SHA). On the other hand, two covariates were negatively associated with the drug-trying behaviour of the students: family's attitude towards drinking (AlPar1) and gender (Gender). These findings of the latent class regression model were discussed in details in Section 7.2.8.

Though the latent class analysis provided a plausible classification of the students based on their drug-trying patterns for this research, in order to provide an alternative means of classification of the students, K-means clustering was conducted, which supported the classification made by the latent class analysis: class 1 was associated with clusters 1 and 2; class 2 was associated with cluster 3, and class 3 was associated with cluster 4.

Chapter 8

Conclusion

Drug use problem is a global issue and has presented a long-term problem in the United Kingdom. Over the years, the United Kingdom Government has devoted its efforts to devising policies aimed at combating drug use problem in the country. To provide helpful guidance to the United Kingdom Government in developing its drug policies, researchers have continuously conducted related drug use studies in order to understand drug-trying behaviour of young people and to explore factors that were associated with such drug-trying behaviour. The "Smoking, Drinking and Drug Use among Young People in England" survey series is a major and exemplary annual survey series in England. However, a review of this survey series revealed several limitations, in particular, the extent of the data analysis and the handling of non-responses.

This research reviewed the "Smoking, Drinking and Drug Use among Young People in England" 2010 survey (the Year 2010 Survey) in terms of its data collection, data processing and data analysis. The primary research aim was to enrich understanding of young people's drug-trying behaviour in England and factors that were associated with their behaviour and hence to improve the quality of future drug-related studies, through appropriate handling of missing data and, built upon the work done in the Year 2010 Survey, developing and applying

new statistical methodologies to permit analysis of multivariate categorical data collected by the Year 2010 Survey study.

To achieve the aim of this research, the main work done in this research was as follows:

(1) The original data set of the Year 2010 Survey was modified into a cleaner working data set which was more suitable for this research. Modification works included a proper recording of the missing data, combining several variables into a single variable, where appropriate, and collapsing factor levels of eight variables in the original data set. Details of modification of the Year 2010 Survey data set to form the working data set of this research were reported in Chapter 2.

(2) Exploratory data analysis in respect of the working data set was conducted. The purposes were to clearly identify any student who had tried a certain drug from those who had never engaged in such drugs before, which helped deeper understanding of how the behaviours of trying drugs were associated mutually (i.e. drug associations), as well as how the smoking, drinking and drug-related socio-demographic factors were associated with students' drug-trying behaviour. Details of the exploratory data analysis of the working data set were reported in Chapter 3.

(3) Missing data problem was another limitation of the Year 2010 Survey. To properly deal with the missing data problem that existed in the working data set, the following procedures were carried out. Firstly we determined the type of missingness for each variable included in the working data set with explanations and whether the missingness was ignorable. Secondly, we adopted several imputation methods on the working data set and compared the results

of the imputed data sets in respect of different imputation methods, in order to evaluate the differences in parameter estimates. Finally, for the 15 drug-trying response variables in the working data set as well as other smoking, drinking and drug-related socio-demographic covariates, we imputed the missing groups by multiple imputation by chained equations (MICE). As such, missing data could be assigned more unbiased values based on other covariates. Details about handling missing data in the working data set were discussed in Chapter 4.

(4) In addition to the exploratory analysis in Chapter 3, development and application of advanced statistical methodologies are carried out to analyse the working data set. The objectives of this analysis were: (1) to further investigate drug associations, and (2) to further explore the specific contributions of the smoking, drinking and drug-related socio-demographic factors to students' drug-trying behaviour in the Year 2010 Survey. These statistical methodologies conducted in this research were:

(a) Both univariate logistic regression models and log-linear analysis models were applied to the working data set to further explore possible interactions among drug-trying response variables, and to further study the associations of the smoking, drinking and drug-related socio-demographic covariates with the students' drug-trying behaviour. Results of the univariate logistic regression models reported the one-way interactions among the 15 drug-trying response variables as well as that numerous smoking, drinking and drug-related socio-demographic covariates were associated with the students' drug-trying behaviour. Results of the log-linear analysis models reported the two-way interactions among the fifteen drug-trying response variables. Details of both univariate logistic analysis and log-linear analysis including their results were reported in Chapter 5.

(b) A two-parameter item response theory model was implemented on the working data set through two approaches, namely the marginal approach and the fully Bayesian approach, to further investigate the relationships between drug-trying response variables and the students' drug-trying behaviour. Results of the item response theory models reported in Chapter 6 permitted an investigation of the probability of the students in trying each drug and the discrimination among the students within each drug-trying response variable.

(c) A latent class analysis model and K-means clustering model were applied to the working data set to examine the allocation of the students to a specific number of classes, their drug-trying patterns, as well as the pattern of drug-trying behaviour in each class. The latent class analysis aimed at how the students should be best classified in accordance with their drug-trying patterns that might influence the subsequent investigation of these students in this research. In the K-means clustering, the best clustering criteria were identified by observing mean values of the fifteen drug-trying response variables, without any latent variable. In addition, the latent class analysis was combined with the logistic regression model to form a latent class regression model, which explained the relationships between the students of different classes with respect to their drug-trying behaviour and the smoking, drinking and drug-related socio-demographic factors via a latent variable. Details of latent class analysis and K-means clustering as well as their findings were reported in Chapter 7.

8.1 Data Processing

In this research, to reduce the complexity of the original data set of the Year 2010 Survey, efforts were spent to eliminate some unnecessary variables and factor levels as well as to reduce the number of types of missingness from three to one. In summary, three types of modification were made to the original data set.

Firstly, there were an excessive number of variables in the original data set, and a few variables contain excessive number of factor levels. In this research, the working data set was formed by including all drug-trying response variables and selected variables that were closely related to drug-trying among adolescents. Several selected variables were combined with others to form new variables. Within several selected variables, factor levels with similar log odds were combined together. The result was a working data set containing fewer but more relevant variables and several variables were with fewer factor levels.

Secondly, regarding the missing data of the original data set, the challenge was to mitigate the number of missing data groups from three to one. This was done by examining the questionnaire questions and deciding how the following missing data categories should be treated: (1) missingness due to questionnaire design; (2) missingness due to repetitive questions and (3) missingness due to non-response or "I don't know" response.

Further processing of the original data set of the Year 2010 Survey was done to generate the working data set of this research that could derive benefits in three dimensions, namely, (1) increasing the response rate of several variables by assigning missingness by design to appropriate values; (2) reducing the occurrence of empty cells, and (3) reducing the complexity of statistical modelling.

8.2 Findings of Exploratory Data Analysis

Before carrying out further statistical analyses for the purposes of this research exploratory data analysis was conducted on the working data set to explore further the associations and relationships among drug-trying response variables and covariates (i.e. smoking, drinking and drug-related socio-demographic

variables). Compared with the findings in the Year 2010 Survey report, key findings of the exploratory data analysis which supported the findings in the 2010 Year Survey report were:

(1) Results of the percentage contingency table of drug-trying response variables confirmed that cannabis was the most tried drug (9.06 %) by the students, followed by gas (8.09 %). The least tried drug was tranquillisers (0.44 %).

(2) Results of the percentage contingency tables, box plots and polychoric correlation plots consistently showed a strong positive association between smoking and drug-trying behaviour of the students in England. However, there were different patterns of pairwise associations between the smoking variables and the 15 drugs.

(3) Similar to the smoking variable, results of the percentage contingency tables, box plots and polychoric correlation plots were consistent to the finding in the Year 2010 Survey report that there was a positive association between drinking alcohol and drug-trying behaviour of the students in England. Also, there were different patterns of pairwise associations between the drinking variables and the 15 drugs.

(4) For the drug-related socio-demographic variables, results of the percentage contingency tables, box plots and polychoric correlation plots supported the findings in the Year 2010 Survey report that the drug-related socio-demographic variables, namely, (a) age of the students (Age), (b) how often the students had been excluded from schools (ExClAN1) and (c) how often the students played truant (Truant1), were strongly and positively associated with drug-trying response variables. However, these three drug-related socio-demographic variables exerted different patterns of pairwise associations with the 15 drugs.

Moreover, additional key findings in relation to the associations and relationships among drug-trying response variables and covariates (i.e. the smoking, drinking and drug-related socio-demographic variables), which were not reported in the Year 2010 Survey report, were found by the exploratory data analysis. They were:

(1) Results of the percentage contingency tables, box plots and polychoric correlation plots generally showed that the 15 drug-trying response variables were highly correlated with each other.

(2) Results of the percentage contingency tables, box plots and polychoric correlation plots further revealed that the strong positive associations between smoking and drug-trying behaviour of the students in England were highly contributed by the following smoking covariates: (a) the attitude of the students' family towards smoking (CgFam1); (b) the students' cigarette smoking status (CgStat1); (c) number of cigarettes smoked by the students in the previous week (Cg7Num); (d) the frequency of purchasing cigarettes from shops by the students (CgBuyF1); (e) sources of obtaining cigarettes by the students (CgGet); (f) whether there were smokers inside the students' houses (CgWho1) as well as (g) the proportion of people a student knows who smoke (CgEstim).

(3) Results of the percentage contingency tables, box plots and polychoric correlation plots also further revealed that the positive associations between drinking and drug-trying behaviour of the students in England was mainly contributed by the following drinking covariates: (a) the attitude of the students' family towards drinking alcohol (AlPar1); (b) usual frequency of drinking alcohol by the students (AlFreq2); (c) sources of buying alcohol by the students (AlBuy); (d) whether there were drinkers inside the students' houses (AlWho1); (e) types

of incidences when the students drank alcohol (Al4W1) as well as (f) the proportion of people a student knows who drank alcohol (AlEstim).

(4) The three drug-related socio-demographic variables, namely (a) age of the students (Age); (b) how often the students had been excluded from schools (ExClAN1) and (c) how often the students played truant (Truant1), were particularly strongly correlated with the five drugs: cannabis, poppers, cocaine, ecstasy and magic mushrooms.

(5) The Year 2010 Year Survey report stated that "girls were less likely than boys to have taken drugs in the last year". According to the percentage contingency table in respect the gender variable (Gender), it was revealed that the aforesaid statement was valid for 7 drugs (cannabis, magic mushrooms, crack, LSD, ketamine, anabolic steroids and tranquillisers) of which the proportion percentages of male students trying them were slightly higher than female students. On the other hand, for the other 8 drugs (heroin, cocaine, methadone, ecstasy, amphetamines, poppers, gas and other drugs), the results of the percentage tabulates showed the opposite. Similarly, the Year 2010 Survey report stated that the school-level variable (percentage of pupils eligible for the free school meals) was not significantly associated with drug use in the survey. However, results of the percentage contingency table in respect whether students have enrolled in free school meal scheme (FSM1) indicated that students involved in the free school meal scheme were more likely to try cannabis, heroin, cocaine, magic mushrooms, methadone, ketamine, gas and tranquillisers.

The above additional key findings by the exploratory data analysis in relation to the associations and relationships among drug-trying response variables and covariates reflected that the data analysis could be further enhanced by employing more sophisticated statistical models to estimate the dependencies between

drug-trying response variables and other related covariates as well as to further study the relationships between drug-trying response variables.

8.3 Multiple Imputation

Another challenge to this research was to manage the missing data in the working data set with appropriate values, such that statistical inferences could be properly interpreted. In the working data set for this research, there was on average approximately 4 % of the data missing in each variable, with a range between 0.58 % and 16.98 %. If the missing data were not imputed properly, bias on estimates might occur. In order to overcome potential limitations caused by missing data, this research successfully utilised a fully Bayesian framework and also multiple imputation by chained equations (MICE). Details of employment of fully Bayesian framework and MICE scheme were discussed in Chapter 4. As discussed in Chapter 4, the missingness of the working data set was diagnosed as MAR by both 'Little test' and the Ridout and Diggle (1991) test. If the data missingness is MAR, then one can refer the missingness as ignorable. In the previous survey work carried out by Fuller et al. (2011) team, the assumption of the missingness being ignorable was also made. Combining the results of the working data set as MAR and ignorable, we were able to impute the missing data by either the MICE scheme or under fully Bayesian framework. The fully Bayesian framework has the advantage of being a one-stage method, when compared to the two-stage method of the MICE scheme. However, the coding of the missingness model can be very complex under the fully Bayesian framework. Under MICE scheme, logistic regression, including polynomial logistic regression, was employed to model nominal variables, whereas normal regression was used to model numerical variables. The regression models that were adopted for imputing the missing values of each variable were conditioned on all other variables. As such, every variable was fitted with an appropriate imputation

model, and the resulting working data sets were fitted with a substantive model. In contrast, in the fully Bayesian framework, a substantive model was fitted on drug-trying response variables. In the fully Bayesian framework, because there were no covariates with missingness, the imputation model was the same as the substantive model.

In this research, the working data set was assumed to be missing at random and generally the MICE scheme was adopted for imputing all variables that contained missingness. Rubin's rule with Wald's test was adopted to test the significance of a covariance or an interaction term in the corresponding regression models employed.

8.4 Findings From Further Investigation of Associations Among Drug-trying Response Variables

Following the additional finding from exploratory data analysis that the 15 drug-trying response variables were highly correlated with each other, in this research, advanced statistical methodologies were needed to further investigate and explore how the 15 drug-trying response variables were associated with each other as well as the extent of their associations. For such purposes, we developed and applied several statistical methodologies, namely, univariate logistic regression models, log-linear analysis models and item response theory models, to the working data set.

Tables 8.4.1 and 8.4.2 below present summary and comparison of key findings in respect of associations among 15 drug-trying response variables from the Year 2010 Survey and various statistical methodologies in this study.

Table 8.4.1: Summary of Key Findings i.r.o. Associations Among 15 Drug-trying Response Variables (Table 1)

	The Year 2010 Survey	This study			
Statistical Methodology	**Logistic regression model** - Fitting a binary drug response (tried any drug or not) with covariates	**Exploratory data analysis** - Percentage contingency tables. - Box plots. - Polychoric correlation plots.	**Univariate logistic regression models** - Fitting each drug with other 14 drugs. - Backward eliminations by Rubin's rule with Wald's test. - Log-odds ratio heat plot. - Covariate sign plot.	**Log-linear analysis models** - Fitting all 15 drugs simultaneously. - Backward eliminations by Rubin's rule with Wald's test. - Log-odds ratio heat plot. - Covariate sign plot.	**Item response theory models** - Two-parameter item response theory models. - Discrimination and difficulty factors for every drugs were considered.
Similar Key Findings	(Associations among 15 drugs were not fully explored) - Cannabis was the most widely used drug.	- Cannabis was the most tried drug, followed by gas. The least tried drug was tranquillisers.		- Cannabis and gas were two drugs with higher proportions of students trying them.	- The highest proportion of students was found to have tried cannabis. - Relatively lower proportions of students were found to have tried tranquillisers, anabolic steroids and other drugs.

Table 8.4.2: Summary of Key Findings i.r.o. Associations Among 15 Drug-trying Response Variables (Table 1 continued)

	The Year 2010 Survey	This study			
Additional Key Findings	- The prevalence of drug use among young people has declined over the period from 2001 to 2010.	- The 15 drugs were highly correlated with each other. - A sustained prevalence of drug use among young people in England.	- Almost all 15 drugs were positively associated with each other drug. - The extent of associations among drugs varied among 15 drugs. For examples, cannabis was found positively associated with 10 drugs, cocaine with 7 drugs and crack with 4 drugs.	- Large number of significant interaction terms among all 15 drugs were found. This supported the found associations among 15 drugs by univariate logistic regression models. - Cannabis was the most dominant drug that has the greater number of significant interaction terms with other types of drugs.	- 6 drugs (ecstasy, cocaine, amphetamines, LSD, heroin and magic mushrooms) were found to exert higher influences on students' drug-trying behaviour. - 3 drugs (anabolic steroids, gas and other drugs) were found to exert relatively lesser influences on students' drug-trying behaviour.
Reference for details	Chapter 1 S. 1.3.3	Chapter 3 S. 3.1.4 and 3.2	Chapter 5 S. 5.2.4	Chapter 5 S. 5.3.4	Chapter 6 S. 6.4 and 6.5

Associations among 15 drug-trying response variables were not fully reported in the Year 2010 Survey report and the key finding in this aspect from the exploratory data analysis was summarised in Section 8.2. The rest of this section focuses on the additional key findings generated from the univariate logistic regression models, log-linear analysis models and item response theory models.

15 univariate logistic regression models were fitted to each drug-trying response variable. Backward eliminations by Rubin's rule with Wald's test were adopted within the univariate logistic regression models, in order to discard insignificant terms. In addition, a log-linear analysis model was fitted to all 15 drug-trying response variables simultaneously to investigate two-way interactions of drug-trying response variables. A backward elimination by Rubin's rule with Wald's test was also adopted to discard insignificant terms. The aim of adopting backward elimination was to identify important parameters and terms and focus on interpreting and elaborating on them. Details of application of both univariate logistic regression models and log-linear analysis models to investigate the relationship and association among the 15 drug-trying response variables could be referred to Chapter 5.

In terms of the numerical results of the univariate logistic regression final models in respect of 15 drug-trying response variables, it was found as a general picture that almost all the 15 drugs were positively associated with each other indicating that if a student has tried a specific drug, the student was more likely to try the other drugs. According to MICE scheme 2, cannabis was found to have positive associations with 10 other drugs. Cocaine, magic mushrooms and ecstasy were found to have positive associations with seven other drugs. Poppers, amphetamines and gas were found to have positive associations with six other drugs. Heroin, tranquillisers and anabolic steroids were found to have positive associations with five other drugs as well as LSD and ketamine, but the

latter two included a negative interaction relationship with one drug. Finally, methadone, crack and other drugs were found to have positive associations with four other drugs. The extent of interaction relationships among drug-trying response variables varied among the 15 drugs. Detailed discussion on the specific interaction relationships among the 15 drugs could be referred to Section 5.2.4.

The results from the log-linear analysis models were found generally comparable with the univariate logistic regression models, particularly in the following two dimensions:

(1) A large number of significant interaction terms among all drugs, in terms of log-odds ratios, were found, and most of these interaction terms were positive, with only a few being negative.

(2) Among the 15 drugs, cannabis was the dominant drug that yielded the greatest number of significant interaction terms with other types of drugs.

Detailed discussion on the significant interaction terms among the 15 drugs presented in the log-linear analysis models could be referred to Section 5.3.4.

When compared the univariate logistic regression saturated models with the log-linear analysis saturated models, the univariate logistic regression saturated models yield less negative terms than the log-linear analysis saturated models. Both models contain mostly positive terms, though the coefficients for the univariate logistic regression saturated models are generally smaller than those for the log-linear analysis saturated models.

We also developed and applied the item response theory models to the drug-trying response variables, in order to discover each student's propensity of trying

drugs. This also included variation among the students in trying every drug and the likelihood for the students to try every drug. The factor scores were used to measure the propensity of each student to try every drug. Discrimination and difficulty factors for every drug were used to measure the influence of each drug to the overall drug-trying behaviour of students and the proportion of young people that tried each drug respectively. The greater discrimination factor coefficient indicated the greater influence of the drug on the overall drug-trying behaviour of students. Similarly, the greater difficulty factor coefficient indicated the smaller proportion of young people who tried a drug and vice versa. In this research, two-parameter item response theory model was implemented on the working data set through two approaches, namely the marginal approach and the fully Bayesian approach, to further investigate the relationships between drug-trying response variables and the students' drug-trying behaviour. Details of application of the two-parameter item response theory models under marginal approach and fully Bayesian approach on the working data set could be referred to Chapter 6.

In all two-parameter item response theory models under marginal approach and fully Bayesian approach, the estimates of the discrimination factors consistently showed that ecstasy, cocaine, amphetamines, LSD, heroin and magic mushrooms were ranked the top six drugs in terms of their high mean estimate values with ecstasy yielded the highest mean estimate value. On the other hand, other drugs, anabolic steroids and gas were consistently ranked the bottom three drugs in terms of their low mean estimate values with gas yielded the lowest mean estimate value of around 1. The aforesaid findings shed additional light on the relationships between drug-trying response variables and the students' drug-trying behaviour. Six drugs, namely ecstasy, cocaine, amphetamines, LSD, heroin and magic mushrooms, were found to exert higher influences on the students' drug trying behaviour that for example, if a student has tried ecstasy,

there was a higher likelihood that the student will try other types of drug.

Also, in all the two-parameter item response theory models under marginal approach and fully Bayesian approach, the estimates of the difficulty factors of all the 15 drug-trying response variables were found to be greater than 1.5, with the majority found to be between 2.5 and 3.2. This observation generally reflected the low proportion of the students who had ever tried each of the 15 drugs. However, amongst the 15 drugs, the estimates of the difficulty factors consistently showed that cannabis, poppers, cocaine, magic mushrooms, ecstasy and amphetamines have relative lower mean estimate values with cannabis yielded the lowest mean estimate value. This reflected that there was the highest proportion of students who had tried cannabis. On the other hand, tranquillisers, anabolic steroids and other drugs were consistently found to have relatively higher mean estimate values with anabolic steroids yielded the highest mean estimate value. Detailed discussion on the results of the two-parameter item response theory models under marginal approach and fully Bayesian approach could be referred to Sections 6.4 and 6.5 respectively.

Overall, findings from the univariate logistic regression models, log-linear analysis models and two-parameter item response theory models consistently supported and explained that there were high correlations among the 15 drug-trying response variables and that each drug exerted different extent of influences on the students' drug-trying behaviour. These findings enrich understanding on the drug-trying behaviour of young people in England in terms of a deeper understanding of the interactions among the 15 drugs, which is one of the objectives of this research. For example, with the finding that cannabis was the most dominant drug that positively associated with 12 other drugs, cannabis can be a good predictor of trying other drugs by young people in England.

8.5 Findings From Further Investigation of Associations Between Drug-trying Response Variables and the Smoking, Drinking and Drug-related Socio-demographic Covariates

Another objective of this research is to identify and understand the factors that are associated with the students' drug-trying behaviour. The Year 2010 Survey has reported that the factors of age, sex, ethnicity, smoking, drinking alcohol, truancy and exclusion were found significantly associated with drug use among the students, albeit in different directions (Fuller et al., 2011). On the other hand, the findings of the exploratory data analysis on the working data set of this research not only supported the aforesaid findings in the 2010 Year Survey report but also provided additional statistical information on how the smoking, drinking and drug-related socio-demographic covariates associated with students' drug-trying behaviour. Considering the additional findings of the exploratory data analysis, advanced statistical methodologies were needed to further investigate what and how were the smoking, drinking and drug-related socio-demographic covariates associated with the students' drug-trying behaviour. For such purposes, we developed and applied several statistical methodologies, namely, univariate logistic regression models, latent class regression models and K-means clustering, to the working data set.

Tables 8.5.1 to 8.5.5 below present summary and comparison of the key findings in respect of associations between drug-trying response variables and the smoking, drinking and drug-related socio-demographic covariates from the Year 2010 Survey and various statistical methodologies in this study.

Table 8.5.1: Summary of Key Findings i.r.o. Associations Between Drug-trying Response Variables and the Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 1)

	The Year 2010 Survey	This study			
Statistical Methodology	**Logistic regression model** - Fitting a binary drug response (tried any drug or not) with covariates	**Exploratory data analysis** - Percentage contingency tables. - Box plots. - Polychoric correlation plots.	**Univariate logistic regression models** - Fitting each drug with other 14 drugs, smoking, drinking and socio-demographic covariates. - Backward eliminations by Rubin's rule with Wald's test. - Covariate sign plot.	**Latent class analysis models** - Classified students by types of drugs tried (3 classes). - Latent class regression model. - Backward eliminations by Rubin's rule with Wald's test.	**K-means clustering model** - Alternative algorithm of latent class analysis. - Divided students by types of drugs tried (4 clusters).

Table 8.5.2: Summary of Key Findings i.r.o. Associations Between Drug-trying Response Variables and the Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 1 continued)

Similar Key Findings	- Age of students was positively associated with drug use. - Smoking and drinking were significantly associated with drug use. - Gender of students was associated with drug use with a slightly higher proportion of male students than female students reported taken drugs. - Truancy and exclusion from schools contributed to general drug use.	- Smoking and drinking were positively associated with drug use. - Age of students was strongly and positively associated with drug use. - Gender of students was associated with drug use. More male students tried 7 drugs (cannabis, magic mushrooms, crack, LSD, ketamine, anabolic steroids and tranquillisers) but more female students tried other 8 drugs (heroin, cocaine, methadone, ecstasy, amphetamines, poppers, gas and other drugs). - Truancy and exclusion from schools were strongly and positively associated with drug use.	- Numerous smoking and drinking covariates explained students' drug-trying behaviour in different extent. - Age, gender, truancy and exclusion from schools explained students' drug-trying behaviour in different extent.	- Some smoking and drinking covariates were positively associated with students' drug-trying behaviour. - Age and truancy were positively associated with students' drug-trying behaviour.	

Table 8.5.3: Summary of Key Findings i.r.o. Associations Between Drug-trying Response Variables and the Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 2)

	The Year 2010 Survey	This study			
Statistical Methodology	**Logistic regression model** - Fitting a binary drug response (tried any drug or not) with covariates	**Exploratory data analysis** - Percentage contingency tables. - Box plots. - Polychoric correlation plots.	**Univariate logistic regression models** - Fitting each drug with other 14 drugs, smoking, drinking and socio-demographic covariates. - Backward eliminations by Rubin's rule with Wald's test. - Covariate sign plot.	**Latent class analysis models** - Classified students by types of drugs tried (3 classes). - Latent class regression model. - Backward eliminations by Rubin's rule with Wald's test.	**K-means clustering model** - Alternative algorithm of latent class analysis. - Divided students by types of drugs tried (4 clusters).

Table 8.5.4: Summary of Key Findings i.r.o. Associations Between Drug-trying Response Variables and the Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 2 continued)

Additional Key Findings	- Ethnicity was associated with drug use. - The most likely sources to obtain helpful information about drugs were: teachers, television and parents.	- Strong positive associations between smoking and drug ise were highly contributed by 7 smoking covariates: CgFam1, CgStat1, Cg7Num, CgBuyF1, CgGet, CgWho1 and CgEstim - Positive associations between drinking and drug use were mainly contributed by 6 drinking covariates: AlPar1, AlFreq2, AlBuy, AlWho1, Al4W1 and AlEstim. - Age, truancy and exclusion from schools were particularly strongly correlated with 5 drugs: cannabis, poppers, cocaine, ecstasy and magic mushrooms.	- The smoking covariates which explained students' drug-trying behaviour included: CgFam1, CgStat, CgGet, CgWho1, CgPe1 and CgIn1. - The drinking covariates which explained students' drug-trying behaviour included: AlFreq, AlEstim, AlBnPub, AlPar1, AlBuy1, AlBuy2, AlBuy, AlUs1, AlUs2, AlPe1, AlIn1 and Al4W1. - Other drug-related socio-demographic covariates which explained students' drug-trying behaviour included: DgPe1, DgIn1, DgEstim, Books1, FSM1 and SHA.	- The smoking covariates which were positively associated with students' drug-trying behaviour included: CgStat1, CgPk1 and CgBuyF1. - The drinking covariates which were positively associated with students' drug-trying behaviour included: AlFreq2 and Al4W1. AlPar1 was found negatively associated with students' drug-trying behaviour.	- Results of the K-means clustering model supported the validity of classification of the students in the latent class analysis models.

Table 8.5.5: Summary of Key Findings i.r.o. Associations Between Drug-trying Response Variables and the Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 2 continued)

Additional Key Findings (cont'd)		- Students involved in free school meal scheme (FSM1) were more likely to try 8 drugs: cannabis, heroin, cocaine, magic mushrooms, methadone, ketamine, gas and tranquillisers. - Students obtained information about drugs from professionals and police were apparently less likely to try drugs.		- Other drug-related socio-demographic covariates which were positively associated with students' drug-trying behaviour included: DgEstim and SHA. Gender was found negatively associated with students' drug-trying behaviour.	
Reference for details	Chapter 1: S.1.3.3	Chapter 3: S.3.3	Chapter 5: S.5.2.5	Chapter 7: S.7.2.8	Chapter 7: S.7.4

An elaborated summary of similar and different (additional) findings of the Year 2010 Survey and the exploratory data analysis in this study could be referred to Section 8.2. The rest of this section focuses on the comparison of key findings from the univariate logistic regression models, latent class analysis models and K-means clustering model.

Details of the application of the univariate logistic regression models to investigate the relationship and association between the 15 drug-trying response variables and the smoking, drinking and drug-related socio-demographic covariates could be referred to Section 5.2.5. In each of the univariate logistic regression models, all the smoking, drinking, drug-related socio-demographic covariates as well as other drug-trying responses were included in the saturated model. Backward elimination was adopted to discard insignificant variables. The univariate logistic regression models provided a more detailed perspective of how trying each drug was caused by other factors. Generally, from the results of the univariate logistic regression models, students' behaviour of trying various drugs could be explained by numerous smoking, drinking and drug-related socio-demographic covariates in different extent. These covariates replaced several drug covariates in predicting whether a student had ever tried at least one of the 15 drugs. These covariates were summarised as follows:

Smoking covariates included: (1) family attitudes toward smoking; (2) cigarette smoking status; (3) sources of purchasing cigarettes; (4) number of smokers who the students know and where those smokers smoked and (5) education and information about smoking.

Drinking covariates included: (1) time and frequency of consuming alcohol; (2) number of alcohol drinkers students know and where those drinkers drank; (3) family's attitude towards drinking; (4) how students purchase alcohol and

where they consume the alcohol; (5) having lessons or obtaining information about drinking and (6) types of issues happened when a student drank alcohol.

Drug-related socio-demographic covariates included: (1) having lessons or obtaining information about drugs; (2) number of smokers students know and where those drug-takers tried drugs; (3) the amount of books students possessed; (4) age; (5) gender; (6) free school meal scheme; (7) frequency of truancy; (8) frequency of being excluded and (9) Strategic Health Authority (SHA).

Detailed discussion on the results of the univariate logistic regression models with drug-trying response variables and covariates could be referred to Section 5.2.5.1.

The latent class analysis models contributed to this research, in addition to the univariate logistic regression models, by providing separate covariate estimate set for each classified group based on drug types. The latent class regression model involved two stages. The first stage involved fitting a latent class analysis model without covariates. It was then followed by retrieving the class probabilities for all students and fitting a Dirichlet distribution regression model on all smoking, drinking, and drug-related socio-demographic variables by backward elimination to select covariates that were related to drug use among adolescents. The second stage involved selecting covariates with Rubin's rule based on the results of both Dirichlet distribution regression model and logistic regression models. In this research, 22 smoking, drinking and drug-related socio-demographic covariates were selected and included into the latent class regression model which explained the relationship between the students of different classes with respect to their drug-trying behaviour and the smoking, drinking and drug-related socio-demographic factors via a latent variable. The latent class regression model was conducted by backward elimination by Rubin's

rule with Wald's test. Details of the application of latent class analysis models and latent class regression models to the working data set could be referred to Section 7.2.

Results of the final latent class regression model revealed that nine covariates were positively associated with the drug-trying behaviour of the students. These nine covariates were: frequency of smoking (CgStat1), type of smoking (CgPk1), frequency of purchasing cigarette (CgBuyF1), frequency of drinking (AlFreq2), reason of drinking (Al4W1), the proportion estimate of drug-takers (DgEstim), frequency of truant (TruantN), age (Age) and Strategic Health Authority (SHA). On the other hand, two covariates were negatively associated with the drug-trying behaviour of the students: family's attitude towards drinking (AlPar1) and gender (Gender). The detailed discussion of the findings of the latent class regression model could be referred to Section 7.2.8.

K-means clustering was an alternative algorithm of latent class analysis in stratifying students based on the pattern of drug responses, without connection to other covariates. Though the latent class analysis models provided a sensible classification of the students based on their drug-trying behaviour, for this research, in order to provide another perspective of classification of the students, K-means clustering was conducted and discussed in Section 7.3. In general, K-means clustering supported the classification made by the latent class analysis models.

Overall, findings from the univariate logistic regression models and latent class analysis models supported findings of the Year 2010 Survey that smoking, drinking and some drug-related socio-demographic (e.g. age, truancy and exclusion from schools) covariates were positively associated with the students' drug-trying behaviour. Additional findings from these advanced statis-

tical methodologies further explained how numerous smoking, drinking and drug-related socio-demographic covariates contributed to the students' drug-trying behaviour at different extent. These additional findings thus provide a deeper understanding on the drug-trying behaviours of young people in England in terms of the associations between drug-trying response variables and the smoking, drinking and drug-related socio-demographic covariates.

8.6 New Methodology for Backward Elimination

One of the research objectives was to develop a new methodology to investigate the association among drug-trying response variables. In order to take imputation and integrated selection of class at each step of backward elimination, we developed a new methodology for the backward elimination of latent class analysis models by Rubin's rule. Most latent class analysis models with backward elimination involved determining the optimal number of latent class, then discarding insignificant covariates one by one, but without re-evaluating the optimal number of latent class. Unlike most latent class analysis models with backward elimination, the newly developed latent class analysis models took imputation into account as well as incorporated Rubin's rule with Wald's test into account. The newly developed latent class analysis models with backward elimination provided a more thorough evaluation of the optimal number of latent class and covariate elimination from saturated model. This was because at each step, the optimal number of latent class was determined, followed by discarding the most insignificant covariate. However, there are limitations to this new methodology. Firstly, each step requires intensive computation of latent class regression models. Secondly, for each step, the number of covariates cannot be too small or too large. Too few covariates might lead to fitting problems and too many covariates might lead to the fitting barely progressing or not progressing at all. The detailed description of the new methodology could be

referred to Section 7.2.6.

8.7 Contributions of the Research

This research contributes to empirical research involving data analysis and drug use related research in different dimensions. Major contributions of this research are:

(1) Grounded on the literature that the robustness of the data analysis may be adversely affected if the missingness problem in a data set of an empirical research is not properly managed, this research showed proper ways to deal with missing data, which are ubiquitous in survey data sets, through the employment of three models:

(a) The substantive model which concerns addressing the questions of interest, for example, in this research, finding the factors that attribute to drug-trying behaviour.

(b) The missingness model which is used to diagnose the mechanism by which the data is missing.

(c) The imputation model which formulates the methodology for imputing the data for data analysis.

(2) This research showed how to enhance the quality of data analysis in an empirical research in order to generate more informative findings relevant to the research objectives from a data set. This was done through the employment of various sophisticated statistical methodologies such as univariate logistic analysis model, log-linear analysis model, item response theory model, latent class

analysis regression model and K-means clustering model, where appropriate.

(3) In this research, a new methodology for the backward elimination of latent class analysis models by Rubin's rule was developed. The newly developed latent class analysis models took imputation into account as well as incorporated Rubin's rule with Wald's test into account. The newly developed latent class analysis models with backward elimination provides a more thorough evaluation of the optimal number of latent class and covariate elimination from saturated model.

(4) Relating to drug use research, the findings from various sophisticated statistical models in this research, that the 15 drugs in question have positive associations with each other in different extent and direction, shed additional light on the drug-trying behaviour of young people among the 15 drugs. Such deeper understanding would provide helpful guidance on formulating policies to combat against drug use problem in England. For example, in terms of resources and effort, relatively more should be inserted and devoted in the direction to combat certain types of drugs that deserve higher priority among the 15 drugs in question, such as cannabis and drugs including ecstasy, cocaine, LSD, magic mushrooms and amphetamines. Cannabis was found to be the most popular and dominant drugs tried by the students and those drugs, including ecstasy, cocaine, LSD, magic mushrooms and amphetamines, were found to exert higher influences (in terms of trying that drug increase the likelihood of trying other drugs associated with that drug) on the students' drug-trying behaviour.

(5) The findings from univariate logistic regression models and latent class regression models in this research, that numerous smoking, drinking and drug-related socio-demographic factors were significantly associated with the students' dug-trying behaviour in different extent and direction. These findings

contribute to a deeper understanding of the drug use problem in England and add evidence to the drug related research literature in two aspects. On one hand, these findings supported the prior research findings that factors like smoking, drinking, age, truancy and exclusion, were positively associated with the students' drug-trying behaviour. On the other hand, these findings further explained how these smoking, drinking and drug-related socio-demographic factors influenced the students' drug-trying behaviour. For example, the findings in this research that smoking and drinking factors were significantly associated with the students' drug-trying behaviour through their related covariates including frequency of smoking (CgStat1), type of smoking (CgPk1), frequency of purchasing cigarette (CgBuyF1), frequency of drinking (AlFreq2) and reason of drinking (Al4W1) respectively. The aforesaid deeper understanding on the effect of smoking, drinking and drug-related socio-demographic factors on the students' drug-trying behaviour would also provide helpful guidance on formulating policies to deal with drug use problem among young people in England.

8.8 Limitations of the Research

Similar to other research studies, this research is subjected to practical limitations which may restrict achievement of the research objective of enhancement of the quality of data analysis to the highest level through appropriate handling of missing data, developing and applying new statistical methodologies.

8.8.1 Limitations of using Unweighted Data

The selected data source of this research is originated from the Year 2010 Survey study. Similar to other survey data based on samples, generally the precision of sample estimates generated from the survey data source is subject to sampling

errors as well as other sources of inaccuracy including non-response bias and over- and under-reporting. As mentioned in Section 2.1, the Year 2010 Survey was a multi-stage sample design stratified by the 10 Strategic Health Authority (SHA) regions in England and hence selection weights were needed in data analysis. However, we used unweighted data in data analysis for this research with reasons in Section 2.3. The use of unweighted data in data analysis in effect assumed the Year 2010 survey is a simple random sample design despite the facts that: (1) as the populations of the ten SHA regions in England were different, there were unequal selection probabilities for students in the ten SHA regions (Fuller et al., 2011) and (2) the stratified structure of the Year 2010 Survey might imply an adverse "neighbourhood" effect on independence of responses in each SHA region. According to Rafferty (2016), not taking sampling weight of a stratified random sample into account may induce sampling errors on estimates, which may then affect true standard errors of variables. Indeed, sampling errors in a multi-stage sample design are not the same as they would have been for a simple random sample of the same size and this needs to be taken into account when calculating standard error of a variable.

Using unweighted data of a multi-stage sample design may over- or under-report the standard error of a variable depends on the property of the variable. In other words, incorporating sampling weights into data analysis of data from a multi-stage sample may increase or decrease the unweighted standard errors of variables. This can be illustrated by two analyses: (1) true standard errors and design factors for five key variables by gender in the Year 2010 Survey (extracted from Tables B.1 to B.5, Appendix B of the Year 2010 Survey Report (Fuller et al., 2011)) as shown in Table F.1.1 in Appendix F and (2) comparison of results of final univariate logistic regression (with backward elimination) among 15 drug-trying response variables between unweighted and weighted models as shown in Table F.2.1 to Table F.2.8 in Appendix F.

In Table F.1.1, the calculation of the true standard errors and design factors was carried out in Stata using a Taylor Series expansion method (Fuller et al., 2011). Table F.1.1 showed that the design factors of all five key variables by gender were slightly greater than 1 which indicates that the true standard errors of the five key variables' estimates increased slightly after incorporating selection weights in data analysis by the researchers of the Year 2010 Survey.

Tables F.2.1 to F.2.8 showed that the final univariate logistic regression analyses among drug-trying response variables lead to increases in some true standard errors of the estimates as well as decreases in the rest when compared the unweighted model with weighted model. However, the differences between all estimates in the unweighted and weighted models were small as all differences were all within one standard error in either unweighted or weighted models.

The above observations are consistent with Stapleton and Kang (2016) that they indicate minor statistical effects if ignoring sampling weighting in data analysis of this study. Nevertheless, we still cannot deny the fact that there is a mismatch of sample design in our data analysis which is a potential source of bias to the results of our data analysis under various statistical models.

8.8.2 Other Limitations

Other practical limitations include the following:

(1) Relating to managing the missing data, in the imputation process of applying MICE (multiple imputation by chained equations) to the working data set, we should make as fewer assumptions as possible. Hence, if a variable before imputation was either ordinal or continuous, the variable was treated as

a nominal variable. This was because ordinal variables are subject to an extra assumption that the odds of trying a certain drug increased when the variable level increased as well as continuous variables are subject to an extra assumption that the increase in the odds was constant between adjacent levels. Treating an ordinal or a continuous variable as a nominal variable requires the least assumptions to the variable.

(2) Relating to the log-linear analysis, because we intended to compare both saturated and final log-linear analysis models with corresponding saturated and final univariate logistic regression models with drug-trying response variables only, merely two-ways interactions among the 15 drugs were considered. Three or more dimensional interactions among the 15 drugs were excluded from both saturated and final log-linear analysis models.

(3) The latent class regression model employed in this research was subjected to the following limitations:

(a) If the positive response rate is too low, the result may fluctuate wildly, and the estimates may be unstable.

(b) The latent class analysis is computationally extensive. It is computationally impossible to include all the smoking, drinking and drug-related socio-demographic covariates in the latent class analysis with backward elimination. We, therefore, have to pre-select those covariates for the latent class analysis with backward elimination.

(4) The K-means clustering model employed in this research is subjected to the following limitations:

(a) K-means clustering only classifies cases according to the Euclidean distances between individuals and their mean point. For data sets with only binomial response variables, regardless of the drug-trying pattern of a student, all cases would be treated the same, as long as their Euclidean distances are the same.

(b) If we wish to integrate regression analysis into the K-means, a two-stage analysis is required, which may result in loss of data information.

8.9 Further Research Work

This research can be potentially extended in several ways. Firstly, we can extend the item response theory models to allow regression models on the factor scores and the difficulty factors. If we want to further investigate the likelihood of students trying drug, we can conduct a longitudinal study for drug-trying response variables over the survey series. This research is a cross-sectional study, which looks at students' responses at one time. We found that age had a significant contribution in determining drug-trying among the students. It would be beneficial to obtain more details on how the students' drug-trying behaviour evolved over time, which can be investigated through a longitudinal study. For instance, did students use the soft drugs before they began using hard drugs? If so, then there may be an argument for criminalising all soft drug use behaviour. Did particular types of soft drug use lead to hard drug use? These questions are difficult to be answered by merely a cross-sectional study. However, by a longitudinal study, answers to these questions can be discovered. Alternatively, we can apply new statistical methodologies on existing data sets that contain more than one binary variable and covariates, such as data sets of "Smoking, Drinking and Drug Use among Young People in England" surveys of different years. All in all, the aforesaid potential future research work shares the objectives of this research that are: to improve the quality of future drug-related survey study

and to enrich understanding of the smoking, drinking and drug-related socio-demographic factors that were associated with drug use among young people in England.

Bibliography

Agrawal, A. (2006). A latent class analysis of illicit drug abuse/dependence: Results from the national epidemiological survey on alcohol and related conditions. *Addiction*, 102(1):94–104.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, Hoboken, New Jersey, 2nd edition.

Allison, P. D. (2001). *Missing Data*. SAGE Publications, Thousand Oaks, California, 1st edition.

Allison, P. D. (2003). Missing data techniques for structural equation modelling. *Journal of Abnormal Psychology*, 112(4):545–557.

Ames, A. J. (2015). Bayesian model criticism: Prior sensitivity of the posterior predictive checks method. *Dissertation. Greensboro: University of North Carolina*.

Arima, S. (2015). Item selection via bayesian IRT models. *Statistics in Medicine*, 34:487–503.

Baker, F. B. (1961). Empirical comparison of item parameters based on the logistic and normal functions. *Psychometrika*, 36:239–246.

Baker, F. B. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, USA.

Bandeen-roche, K. and Miglioretti, D. L. (1997). Latent variable regression

for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley, Hoboken, New Jersey, USA, 3rd edition.

Bazan, J., Branco, M., and Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis*, 1(4):861–892.

Best, N. and Mason, A. (2012). Bayesian approaches to handling missing data. http://www.bias-project.org.uk/Missing2012/Lectures.pdf. Last accessed on Dec 17, 2017.

Bijleveld, C. C. J. H., van der Kemp, L. J. T., Mooijaart, A., van der Kloot, W. A., van der Leeden, R., and van der Durg, E. (1998). *Longitudinal data analysis: Designs, models and methods*. SAGE Publications Ltd, Thousands Oaks, CA, USA, 1st edition.

Birnbaum, A. (1967). Statistical theory for logistic mental test models with a prior distribution of ability. *Research Bulletin* No. RB-67-12.

Brand, J. P. L. (1999). Development, implementation and evaluation of multiple imputaiton strategies for the statistical analysis of incomplete data sets. *Dissertation. Rotterdam: Erasmus University*.

Buuren, S. V. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242.

Buuren, S. V. (2012). *Flexible Imputation of Missing Data*. CRC Press, Florida, US.

Buuren, S. V. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate impuation by chained equations in R. *Journal of Statistical Software*, 45(3):1–66.

Carlson, J. E. and von Davier, M. (2013). *Item Response Theory*. Educational Testing Service, Princeton, New Jersey, USA.

Carpenter, J. R. and Kenward, M. G. (2013). *Multiple Imputation and Its Application*. John Wiley and Sons, Hokoben, New Jersey, USA.

Casey, D. (2012). *A Fresh Approach to Drugs: The Final Report of the UK Drug Policy Commission*. UKDPC, UK.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. Springer-verlag, New York, USA.

Clements, F. E. (1954). Use of cluster analysis with anthropological data. *American Anthropologist*, 56(2).

Clogg, C. C. and Eliason, S. R. (1987). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16(1):8–44.

Collins, L. M. and Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis*. John Wiley and Sons, Hoboken, New Jersey.

Copello, A. (2009). *Adult Family Members and Carers of Dependent Drug Users: Prevalence, Social Cost, Resource Savings and Treatment Responses*. United Kingdom Drug Policy Commission, UK.

Copps, J. (2013). *No Quick Fix: Exposing the Depth of Britain's Drug and Alcohol Problem 2013*. New Philanthropy Capital, UK.

Dayton, C. M. and MacReady, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Assocation*, 83(401):173–178.

De Leeuw, E. D. (2001). Reducing missing data in surveys: An overview of methods. *Quality and Quantity*, 35(2):147–160.

De Leeuw, E. D., Hox, J. J., and Dillman, D. A. (2008). *International Handbook of Survey Methodology*. European Association of Methodology/Lawrence Erlbaum Associates, New York, USA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38.

Department of Health (2010). *Healthy Lives, Healthy People: Our Strategy for Public Health in England*. The Stationery Office, London, UK.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley and Sons.

Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Chapman and Hall, London, UK, 3rd edition.

Drasgow, F. (1986). Polychoric and polyserial correlations. In Kotz, S. and Johnson, N., editors, *The Encyclopedia of Statistics*, pages 68–74. John Wiley and Sons, Hokoben, New York.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics*, 21(2):362–375.

European Monitoring Centre for Drugs and Drug Addiction (2012a). "Legal topic overviews: Classification of controlled drugs". http://www.emcdda.europa.eu/html.cfm/index5733EN.html. Last accessed on Aug 31, 2017.

European Monitoring Centre for Drugs and Drug Addiction (2012b). "Legal topic overviews: Classification of controlled drugs". http://www.emcdda.europa.eu/html.cfm/index146601EN.html. Last accessed on Sep 3, 2017.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (1993). *Finding Groups in Data : An Introduction to Cluster Analysis Cluster Analysis*. John Wiley and Sons, New York, USA.

Ferrando, P. J. (1994). Fitting item response models to the EPI-A impulsivity subscale. *Educational and Psychological Measurement*, 54:118–127.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression in R. *Journal of Applied Statistics*, 31(7):799–815.

Fink, A. G. (2002). *The Survey Handbook*. SAGE, 2nd edition.

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency vs interpretability of classifications. *Biometrics*, 21:768–769.

Fuller, E., Agalioti-Sgompou, V., Christie, S., Fiorini, P., Hawkins, V., Hinchliffe, S., Lepps, H., Sal, N., and Sharman, S. (2015). *Report of Smoking, Drinking and Drug Use Among Young People in England in 2014*. National Centre for Social Research, UK.

Fuller, E., Bridges, S., Gill, V., Omole, T., Sutton, R., and Wright, V. (2011). *Report of Smoking, Drinking and Drug Use Among Young People in England in 2010*. National Centre for Social Research, UK.

Fuller, E., Gill, V., Hawkins, V., Mandalia, D., and Whalley, R. (2012). *Report of Smoking, Drinking and Drug Use Among Young People in England in 2011*. National Centre for Social Research, UK.

Fuller, E. and Hawkins, V. (2014). *Report of Smoking, Drinking and Drug Use Among Young People in England in 2013*. National Centre for Social Research, UK.

Fuller, E., Henderson, H., Nass, L., Payne, C., Phelps, A., and Ryley, A. (2013). *Report of Smoking, Drinking and Drug Use Among Young People in England in 2012*. National Centre for Social Research, UK.

Gauffin, K. (2013). Childhood socio-economic status, school failure and drug abuse: a swedish national cohort study. *Addiction*, 108:1441 – 1449.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Ghodse, H. (2012). *Ghodse's Drugs and Addictive Behaviour*. Cambridge Press, UK.

Glickman, M. E., Seal, P., and Eisen, S. V. (2009). A non-parametric Bayesian diagnostic for detecting differential item functioning in IRT models. *Health Services and Outcomes Research Methodology*, 9:145–161.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). *Selection Modelling versus Mixture Modelling with Nonignorable Nonresponse. In H. Wainer, Drawing Inferences from Self-Selected Samples*. Springer, New York, USA.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Government, H. (1998). *Tackling Drugs to Build a Better Britain - The Government's Ten-Year Strategy for Tackling Drugs Misuse*. HM Government, London, UK.

Government of the Netherlands (2011). "Difference Between Hard and Soft Drugs". https://www.government.nl/topics/drugs/difference-between-hard-and-soft-drugs. Last accessed on Aug 07, 2017.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):149–192.

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1994). *Multivariate Data Analysis: with readings*. Prentice-hall, Upper Saddle River, New Jersey, USA, 6th edition.

Hale, D. and Viner, R. (2013). Trends in the prevalence of multiple substance use in adolescents in England, 1998-2009. *Journal of Public Health*, 35(3):367–374.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. SAGE, Thousand Oaks, California, USA, 1st edition.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley and Sons, New York, USA.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, 28:100–108.

Hastings, W. K. (1970). Monte Carlo sampling-based methods using markov chains and their applications. *Biometrika*, 57:97–109.

Hazewinkel, M. (1994). *Encyclopaedia of Mathematics*. Springer Netherlands, Netherlands, 1st edition.

Hibell, B. (2011). *2011 ESPAD Report*. European School Survey Project on Alcohol and Other Drugs, Stockholm, Sweden.

HM Government (1971). "Misuse of drugs act 1971, Chapter 38". http://www.legislation.gov.uk/ukpga/1971/38/contents. Last accessed on Aug 07, 2017.

HM Government (2010). *The Coalition: Our Programme for Government*. HM Government, London, UK.

HM Government (2015). *Drug Strategy 2010 'A Balanced Approach' Third Annual Review*. HM Government, London, UK.

HM Government (2017). *2017 Drug Strategy*. HM Government, London, UK.

Hobbs, G. and Vignoles, A. (2007). *Is Free School Meal Status A Valid Proxy for Socio-economic Status (In Schools Research)?* CEE Discussion Papers, Centre for the Economics of Education, LSE, London, UK.

Hsieh, M. and Proctor, T. P. (2010). A comparison of Bayesian MCMC and marginal maximum likelihood methods in estimating the item parameters for the 2PL IRT model. *International Journal of Innovative Management, Information and Production*, 1:1.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of American Statistics Association*, 100:469.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Upper Saddle River, New Jersey, USA.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, 31(3):264–323.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiol*, 64(5):402–406.

Kelejian, H. H. (1969). Missing observations in multivariate regression: Efficiency of a first-order method. *Journal of the American Statistical Association*, 64(328):1609–1616.

Ketchen, D. J. and Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, 17:441–458.

Kyureghian, G., Capps, O., and Nayga, R. M. (2011). *A Missing Variable Imputation Methodology with An Empirical Application, in David M. Drukker (ed.) Missing Data Methods: Cross-sectional Methods and Applications (Advances in Econometrics, Volume 27 Part 1)*. Emerald Group Publishing Limited.

Lader, D. (2015). *Drug Misuse: Findings from the 2014/15 Crime Survey for England and Wales*. Home Office, UK, UK.

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. SAGE Publications, California, US, 1st edition.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. *Measurement and Prediction*, pages 362–412.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. *Journal of American Statistics Association*, 86:1065–1073.

Linzer, D. A. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10).

Little, R. J. (2008). *Selection and Pattern-mixture Models. In Longitudinal Data Analysis, G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs , 409 to 431.* Chapman and Hall, Boca Raton, Florida, USA.

Little, R. J. and Schenker, N. (1995). *Missing Data. In: Arminger G., Clogg C.C., Sobel M.E. (eds) Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Springer, Boston, MA, USA.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Buisness and Economie Statistics*, 6(3):287–296.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134.

Lloyd, S. P. (1982). Least squares quantization in PCM. technical note, bell laboratories. *IEEE Transactions on Information Theory*, 28:128–137.

Loken, E. and Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63:509–525.

Lord, F. M. (1951). *A Theory of Test Scores and Their Relation to the Trait Measured*. Educational Testing Service, Princeton, New Jersey, USA.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three parameter logistic model. *Educational and Psychological Measurement*, 28:989–1020.

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.

Lunn, D. J., Whittaker, J. C., and Best, N. (2006). A bayesian toolkit for genetic association studies. *Genetic Epidemiology*, 30:231–247.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297.

Magidson, J. (1987). Weighted log-linear modeling. *Statistical Association 1987. Proceedings of the Social Statistics Section*, pages 171–174.

Maier, M. J. (2014). Dirichletreg: Dirichlet regression for compositional data in R. *Biometrics*, 47(4):1617–1621.

Mak, T. (1992). Estimation of parameters in heteroscedastic linear models. *Journal of Royal Statistical Society Serie B*, 54(2):649–655.

Manders, B. (2016). *Deaths Related to Drug Poisoning in England and Wales: 2016 Registrations*. International Centre for Drug Policy, UK.

Mathers, N., Fox, N., and Hunn, A. (2007). *Surveys and Questionnaires*. The NIHR RDS for the East Midlands / Yorkshire and the Humber.

May, C. (2017). *Transnational Crime and the Developing World*. Global Financial Integrity, Washington, D.C., USA.

McArdle, P. (2004). Substance abuse by children and young people disease in childhood. *British Medical Journal*, 89:8.

McCullagh, P. and Nelder, J. (1999). *Generalized Linear Models*. Chapman and Hall, USA.

Mckeganey, N. (2004). Preteen children and illegal drugs. *Drugs, Education, Prevention and Policy*, 11(4):315–327.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York, USA.

Metropolis, N., Rosenbluth, A. W., and Rosenbluth, M. N. (1953). Equations of state calculations by fast computing. *Journal of Chemical Physics*, 21:1087–1091.

Mold, A. (2007). Illicit drugs and the rise of epidemiology during the 1960s. *Journal of Epidemiology and Community Health*, 61(4):278–281.

Murphy, K. (2007). Software for graphical models: A review. *International Society for Bayesian Analysis Bulletin*, 14(4):13–15.

Nering, M. L. and Ostini, R. (2010). *Handbook of Polytomous Item Response Theory Models*. Routledge, UK.

Ng, H., Ong, S., Foong, K., Goh, P., and Nowinski, W. (2006). *Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm*. Paper Presented at: SSIAI 2006. Proceedings of the IEEE Southeast Symposium on Image Analysis and Interpretation, Denver, USA.

NHS (2012). *Substance Misuse Among Young People 2011-12*. National Health Service, UK, UK.

Niblett, P. (2016). *Statistics on Drugs Misuse England 2016*. Health and Social Care Information Centre, England, UK.

NTA (2012). *Drug Treatment 2012: Progress Made, Challenges Ahead*. NTA, UK, UK.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. John Wiley and Sons, New Jersey, USA, 1st edition.

Osgood, D. W., McMorris, B. J., and Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitive Criminology*, 18:267–296.

Patz, R. J. and Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24(4):342–366.

Pearson, E. S. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, 30:134–148.

Pharris, A. (2011). Community patterns of stigma towards persons living with HIV: A population-based latent class analysis from rural vietnam. *BMC Public Health*, 11:705.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Rafferty, A. (2016). Session 1: Introduction to complex survey design. URL: https://www.ukdataservice.ac.uk/media/440347/rafferty.pdf.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Educational Testing Service, Copenhagen, Denmark.

Reise, S. P. and Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, 14:45–58.

Ridout, M. S. and Diggle, P. J. (1991). Testing for random dropouts in repeated measurement data. *Biometrics*, 47(4):1617–1621.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5).

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2):259–278.

Rouse, S. V., Finger, M. S., and Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, 72:282–307.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley and Sons, Hoboken, New Jersey, 1st edition.

Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley and Sons, Hoboken, New Jersey, 2nd edition.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Siumlation*, 72(3):217–232.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. D. Reidel, Dordrecht, Austria.

Santini, M. (2016). Advantages and disadvantages of means and hierarchical clustering (unsupervised learning). URL: http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behaviour Research*, 33(4):545–571.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52:333.

Shi, W. and Zeng, W. (2014). Application of K-means clustering to environmental risk zoning of the chemical industrial area. *Frontiers of Environmental Science and Engineering*, 8(1):117–127.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71.

Spiegelhalter, D. J. (2003). *WinBUGS User Manual, Version 1.4*. MRC BioStatistics Unit, Cambridge, UK.

Spiegelhalter, D. J. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28:3049–3067.

Stapleton, L. M. and Kang, Y. (2016). Design effects of multilevel estimates from national probability samples. *Sociological Methods and Research*, 1(28).

Statistics Team, NHS Digital (2017). *Report of Smoking, Drinking and Drug Use Among Young People in England in 2016*. NHS Digital, UK.

Sterne, J. A. C., Kenward, M. G., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, 338.

Stimson, G. V. (1987). British drug policies in the 1980s: a preliminary analysis and suggestions for research. *Addiction*, 82(5):77–488.

Sudman, S. and Bradburn, N. M. (1974). *Response Effects in Surveys*. Aldine, Chicago, USA.

Swiftl, W. (2013). *Analysis of Cannabis Seizures in NSW, Australia: Cannabis Potency and Cannabinoid Profile*. National Drug and Alcohol Research Centre Sydney, Sydney, Australia.

Telgarsky, M. and Vattani, A. (2010). Hartigan's method: K-means clustering without voronoi. *International Conference on Artificial Intelligence and Statistics*, pages 820–827.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological BulletiKorean Journal of Anesthesiol*, 133(5):859–883.

Tu, Y. K. and Greenwood, D. C. (2012). *Modern Methods for Epidemiology*. Springer, New York, USA, 1st edition.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, USA.

United Nations (1961). *Single Convention on Narcotic Drugs, 1961*. United Nations, New York, USA.

United Nations (2016). *Terminology and Information on Drugs*. United Nations Office, Vienna, Austria.

United Nations Office on Drug Control and Crime (2005). *World Drug Report 2005*. United Nations Publication, Vienna, Austria.

United Nations Office on Drug Control and Crime (2015). *World Drug Report 2015*. United Nations Office Publication, Vienna, Austria.

United Nations Office on Drug Control and Crime (2017). *World Drug Report 2017*. United Nations Publication, Vienna, Austria.

Van der Linden, Hambleton, W. J., and K., R. (1997). *Handbook of Modern Item Response Theory*. Springer-Verlag, New York, 1st edition.

Vermunt, J. K. (1996). *Causal Log-linear Modeling with Latent Variables and Missing Data. In U. Engel, and J. Reinecke (Eds.), Analysis of Change: Advanced Techniques in Panel Data Analysis (pp. 35-60)*. Walter de Gruyter, Berlin, Germany.

Vermunt, J. K. (1997). *Log-linear Models for Event Histories*. SAGE Publications, Thousand Oaks, California, 8th edition.

Vermunt, J. K. and Magidson, J. (2005). *Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In A. Van der Ark, M. A. Croon, and K. Sijtsma (Eds.), New Developments in Categorical Data Analysis for the Social and Behavioral Sciences (P. 41-62)*. Erlbaum, Mahwah, New Jersey, USA.

Vermunt, J. K. and Magidson, J. (2007). Latent class analysis with sampling weights: A maximum-likelihood approach. *Sociological Methods and Research*, 36(1).

Vermunt, J. K. and Magidson, J. (2008). *Manual For Latent Gold 4.5 Syntax Module*. Statistical Innovations, Belmont, Massachusetts.

von Davier, M. (2009). *Is There Need for the 3PL Model? Guess What?* Educational Testing Service, Princeton, New Jersey, USA.

Vuolo, M. (2009). National-level drug policy and young people's illicit drug use: A multilevel analysis of the european union. *Drug and Alcohol Dependence*, 131(1):149 – 156.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Weiner, S. P. and Dalessio, A. T. (2006). *Oversurveying: Causes, Consequences, and Cures. A.I. Kraut (Ed.), Getting Action from Organizational Surveys: New Concepts, Methods and Applications (pp. 294-311)*. Ossey-Bass, San Francisco, California, USA.

White, I. R., Royston, P., and Wood, A. M. (2009). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics In Medicine*, 30(4):377–399.

Appendix A

Questions in the Year 2010 Survey Questionnaire and Variables in Working Data Set

A.1 Classification of Questions in the Year 2010 Survey Questionnaire

Table A.1.1: Table of Question Classification of the Year 2010 Survey Questionnaire (Table 1)

Question Number	**Group**
1-6	General question
7-10, 236-237	Smoking frequency
11-22	Attempt to smoking, giving-up smoking and family responses
23-26	General information about purchase of cigarettes
27-37	More in-depth information about purchase of cigarettes
38-39	Smoking with others
40-43	Attempt to drink alcohol and family responses
44-51,55	Habit of drinking alcohol and family responses
52-54, 56-69	Detail about alcohol consumption during the last 7 days
70-89	Detail about alcohol consumption during the last 4 weeks
90-91	Why people of own age drink alcohol
92-99	About cannabis
100-107	About speed, amphetamines
108-114	About LSD
115-121	About ecstasy
122-128	About semeron
129-135	About poppers
136-142	About tranquillisers
143-149	About heroin
150-156	About magic mushrooms
157-163	About methadone
164-170	About crack
171-177	About cocaine
178-184	About ketamine

Table A.1.2: Table of Question Classification of the Year 2010 Survey Questionnaire (Table 2)

Question Number	**Group**
185-191	About anabolic steroids
192-198	About glue, gas, aerosols and solvents
199-206	About other drugs
207-210	General questions of taking drugs but excluding cigarettes or alcohol
211-212	Whether fine for people of same age to take drugs, smoke or drink alcohol
213-216	Number of people of own age who smoke, drink alcohol or take drugs
217-219	Sources of helpful information about smoking, drinking alcohol and taking drugs
220-223	Places of helpful information about smoking, drinking alcohol and taking drugs
224-229	General questions about school
230	Number of books in house
231-235	Number of housemates, number of smoking and drinking housemates
238	Any other questions

A.2 List of Variables in working data set

Table A.2.1: Description of Smoking Variables in Working Data Set

Name	Label
Smoking Variables	
CgFam1	Family attitudes toward smoking
Cg7Num	Average number of cigarettes smoked previous week
CgStat	Cigarette smoking status
CgStat1	Cigarette smoking status (with average number of cigarettes smoked previous week)
CgPk1	Whether usually smoke packet cigarettes, roll-ups or both
CgGet1	Number of sources of buying cigarettes usually through shops/ machine/ Internet
CgGet2	Number of sources of buying cigarettes usually through people
CgGet3	Number of sources of being given cigarettes usually by people or other sources
CgGet	Types of sources of obtaining cigarettes usually
CgPp1	Types of people who know smoke cigarettes
CgWhoHme	Whether people live with smoke inside house
CgWhoSmo	Number of people live with smoke
CgWho1	Smokers in house and where
CgBuyF1	Frequency of buying cigarettes from shop in the past year
CgEstim	How many own age smoke
CgPe1	Getting helpful information about smoking cigarettes from people
CgIn1	Getting helpful information about smoking cigarettes from media
LsSmk	Whether had lessons on smoking in last 12 months
CgNow	Whether smokes cigarettes nowadays

Table A.2.2: Description of Drinking Variables in Working Data Set

Name	Label
Drinking Variables	
AlEvr	Have you ever drunk alcohol
AlFreq	Usual frequency of drinking alcohol
AlLast	When last had alcohol
Al7Day1	How many days in last seven drank alcohol
AlFreq2	Usual frequency of drinking alcohol (with how many days in last seven drank alcohol
AlBnPub	Been in a pub, bar or club in the evening in the last four weeks
AlEstim	How many own age drink?
LsAlc	Whether had lessons on drinking in last 12 months
AlPar1	How do respondent's parents/ guardians feel about drinking alcohol
AlBuy1	Number of places a respondent usually purchase alcohol
AlBuy2	Number of people sources a respondent usually purchase alcohol
AlBuy	Respondents usually purchase alcohol themselves/ through people
AlUs1	Types of people a respondent usually uses alcohol with
AlUs2	Types of places a respondent usually uses alcohol in
Al4W1	Types of issues happening when drinking alcohol in last 4 weeks
AlWhy1	Why do you think people of same age drink?
AlWhoHme	Whether people live with drank inside house
AlWhoDr	Number of people live who drank
AlWho1	Drinkers in house and where
AlPe1	Getting helpful information about drinking from people
AlIn1	Getting helpful information about drinking from media

Table A.2.3: Description of Drug-related Socio-Demographic Variables and Response Variables in Working Data Set

Name	Label
Drug-related Socio-Demographic Variables	
DgPe1	Getting helpful information about drug use from people
DgIn1	Getting helpful information about drug use from media
DgEstim	How many own age take drugs
Books1	How many books in home
LsDrg	Whether had lessons on drug in last 12 months
Age	Age 11 to 15
Gender	Sex of respondents
FSM1	Whether enrolled in free school meal scheme
Truant1	Whether ever truanted
TruantN	How often played truant
ExclA1	Whether ever been excluded
ExclAN1	How often been excluded
SHA	Strategic Health Authority
Response Variables	
DgTdCan1	Ever tried cannabis
DgTdHer1	Ever tried heroin
DgTdCok1	Ever tried cocaine
DgTdMsh1	Ever tried magic mushrooms
DgTdCrk1	Ever tried crack
DgTdMth1	Ever tried methadone
DgTdEcs1	Ever tried ecstasy
DgTdAmp1	Ever tried amphetamines
DgTdLSD1	Ever tried LSD
DgTdPop1	Ever tried poppers
DgTdKet1	Ever tried ketamine
DgTdAna1	Ever tried anabolic steroids
DgTdGas1	Ever tried gas
DgTdOth1	Ever tried other drugs
DgTdTrn1	Ever tried tranquillisers

Appendix B

Tables Related to Univariate Logistic Regression

B.1 Types of Variables Used in Univariate Logistic Regression

Table B.1.1: Table of Types of Variables Used in Logistic Regression Model With Covariates (F: Factor (categorical), L: Linear) (Table 1)

Variable	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack	Methadone	Ecstasy	Amphetamine	LSD	Poppers	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
CgFam1	F	F	F	F	F	F	F	L	L	F	L	F	F	L	L
CgStat1	F	L	F	F	F	L	F	F	F	F	F	L	F	L	L
CgPk1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
CgGet1	F	F	F	F	F	F	F	F	F	F	F	L	F	F	L
CgGet2	F	F	F	F	L	F	F	F	F	F	F	L	L	F	L
CgGet3	F	F	F	F	L	F	F	F	F	F	F	L	F	F	L
CgGet	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
CgPp1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
CgWho1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
CgBuyF1	F	L	F	F	F	F	F	F	F	F	F	L	L	L	F
CgEstim	F	F	L	L	L	F	L	F	L	F	L	F	L	L	L
CgPe1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
CgIn1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
LsSmk	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlLast	F	F	L	L	L	F	L	L	L	L	L	L	L	L	L
Al7Day1	F	L	L	F	L	F	F	F	L	F	F	L	L	L	L
AlFreq2	F	L	F	L	F	F	F	F	F	L	L	L	F	F	L
AlBnPub	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlEstim	F	L	L	L	L	L	L	L	L	L	L	F	L	L	L
LsAlc	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlPar1	L	F	F	L	F	F	L	L	L	L	L	L	F	L	F
AlBuy1	F	F	F	F	L	F	F	F	F	F	F	L	F	L	L
AlBuy2	F	F	F	F	F	F	F	F	F	F	F	F	F	F	L

Table B.1.2: Table of Types of Variables Used in Logistic Regression Model With Covariates (F: Factor (categorical), L: Linear) (Table 2)

Variable	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack	Methadone	Ecstasy	Amphetamine	LSD	Poppers	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
AlBuy	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlUs1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlUs2	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Al4W1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlWhy1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlWho1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlPe1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
AlIn1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
DgPe1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
DgIn1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
DgEstim	F	L	L	L	L	L	F	F	L	F	F	L	F	F	F
Books1	F	L	L	L	L	L	L	L	L	L	L	F	F	L	L
LsDrg	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
age	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
gender	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
FSM1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
TruantN	F	F	F	F	F	L	F	F	F	F	L	F	F	L	L
ExclAN1	F	L	F	L	F	L	F	L	L	F	F	L	F	L	F
SHA	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Cannabis	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Heroin	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Cocaine	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Magic Mushrooms	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Crack	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Methadone	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Ecstasy	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Amphetamine	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
LSD	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Poppers	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Ketamine	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Anabolic Steroids	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Gas	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Other Drugs	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Tranquillisers	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

B.2 Covariates Included in Saturated Models of Logistic Regression with Covariates

Table B.2.1: Table of Covariates Included in Saturated Models of Logistic Regression with Covariates (✓: Included, ✗: Excluded) (Table 1)

Variable	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack	Methadone	Ecstasy	Amphetamine	LSD	Poppers	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
CgFam1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CgStat1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
CgPk1	✗	✗	✗	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓
CgGet1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CgGet2	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓
CgGet3	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗	✓
CgGet	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗
CgPp1	✓	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✗
CgWho1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CgBuyF1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CgEstim	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CgPe1	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗
CgIn1	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
LsSmk	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AlLast	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
Al7Day1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
AlFreq2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
AlBnPub	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AlEstim	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓	✓
LsAlc	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AlPar1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AlBuy1	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓	✓	✗	✓	✓
AlBuy2	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗

Table B.2.2: Table of Covariates Included in Saturated Models of Logistic Regression with Covariates (✓: Included, ✗: Excluded) (Table 2)

Variable	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack	Methadone	Ecstasy	Amphetamine	LSD	Poppers	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
AlBuy	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗	✓	✓
AlUs1	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
AlUs2	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓
Al4W1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
AlWhy1	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓	✓	✗	✗
AlWho1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AlPe1	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
AlIn1	✓	✗	✗	✗	✓	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗
DgPe1	✓	✗	✓	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✓	✗
DgIn1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DgEstim	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Books1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LsDrg	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
age	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
gender	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
FSM1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TruantN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ExclAN1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SHA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cannabis	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Heroin	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cocaine	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Magic Mushrooms	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Crack	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Methadone	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ecstasy	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓
Amphetamine	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
LSD	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Poppers	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
Ketamine	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Anabolic Steroids	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
Gas	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
Other Drugs	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
Tranquillisers	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

B.3 Univariate Logistc Regression Results

B.3.1 Within Response Variables with Backward Elimination

Table B.3.1: Table of Estimates of Univariate Logistic Regression Final Models within Drug-trying Response Variables (Table 1)

Cannabis			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-2.7937 (0.0544)	-2.749 (0.0522)	-2.7497 (0.0524)
Cannabis			
Heroin			
Cocaine	1.9203 (0.4247)	1.8951 (0.4082)	1.7329 (0.3951)
Magic Mushrooms	1.7261 (0.3154)	1.8025 (0.3044)	1.7882 (0.2957)
Crack	1.3056 (0.5562)	1.2621 (0.5125)	1.2098 (0.5254)
Methadone	2.2028 (0.5141)	1.9568 (0.4965)	1.8096 (0.511)
Ecstasy	2.1544 (0.5094)	1.8182 (0.442)	1.6372 (0.4145)
Amphetamines	1.17 (0.5102)	1.2899 (0.4653)	1.067 (0.4404)
LSD	1.9235 (0.6816)		
Poppers	2.951 (0.2274)	3.0107 (0.2211)	2.9916 (0.2161)
Ketamine	3.0345 (0.4917)	2.8491 (0.4756)	2.7259 (0.4577)
Anabolic Steroids	1.6713 (0.5272)	1.5765 (0.4951)	1.4771 (0.5199)
Gas	0.6926 (0.1439)	0.6706 (0.1385)	0.6938 (0.1389)
Other Drugs	1.5111 (0.6388)	1.4385 (0.5998)	
Tranquillisers			
Heroin			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.9025 (0.3635)	-6.6389 (0.315)	-6.2607 (0.2896)
Cannabis			
Heroin			
Cocaine	3.1359 (0.514)	2.8176 (0.5242)	2.6392 (0.621)
Magic Mushrooms	1.1325 (0.5499)		
Crack	2.6261 (0.5796)	2.6312 (0.5713)	2.7598 (0.6416)
Methadone			
Ecstasy			
Amphetamines			
LSD		1.9736 (0.595)	1.3897 (0.68)
Poppers			
Ketamine		1.7981 (0.792)	
Anabolic Steroids	2.0174 (0.7362)		1.9491 (0.8275)
Gas	2.0131 (0.4535)	1.9631 (0.421)	1.6851 (0.3872)
Other Drugs			
Tranquillisers			
Cocaine			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.2757 (0.2793)	-6.152 (0.2603)	-5.978 (0.2611)
Cannabis	2.4658 (0.3672)	2.3927 (0.3573)	2.2457 (0.3506)
Heroin	2.8845 (0.5881)	2.6367 (0.5525)	2.3106 (0.563)
Cocaine			
Magic Mushrooms			
Crack	2.2258 (0.5846)	1.9776 (0.529)	2.0184 (0.5726)
Methadone			
Ecstasy	2.1457 (0.3905)	2.0287 (0.3694)	2.0385 (0.3708)
Amphetamines	1.367 (0.4369)	1.3557 (0.4116)	1.3445 (0.3998)
LSD			
Poppers	1.5163 (0.3495)	1.4097 (0.3362)	1.3961 (0.3672)
Ketamine			
Anabolic Steroids			
Gas			
Other Drugs	1.5313 (0.597)	1.6853 (0.5778)	1.5775 (0.6246)
Tranquillisers			

Table B.3.2: Table of Estimates of Univariate Logistic Regression Final Models within Drug-trying Response Variables (Table 2)

Magic Mushrooms			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-5.5673 (0.1961)	-5.5786 (0.1948)	-5.4764 (0.1918)
Cannabis	2.412 (0.2594)	2.3972 (0.2602)	2.3558 (0.2554)
Heroin	1.3508 (0.5966)	1.7713 (0.5318)	1.4814 (0.5399)
Cocaine			
Magic Mushrooms			
Crack	1.2651 (0.5339)		
Methadone			
Ecstasy		1.0356 (0.383)	0.9847 (0.3881)
Amphetamines	1.8743 (0.3693)	1.6991 (0.3672)	1.5451 (0.4069)
LSD	1.9126 (0.4486)	1.6311 (0.4691)	1.4617 (0.5252)
Poppers			
Ketamine			
Anabolic Steroids			
Gas	1.1088 (0.2611)	1.0917 (0.2565)	1.1226 (0.2515)
Other Drugs	1.68 (0.5097)	1.2422 (0.5485)	
Tranquillisers			1.4506 (0.6681)
Crack			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.6165 (0.3327)	-6.3997 (0.3056)	-6.255 (0.2871)
Cannabis	2.1625 (0.4487)	2.2305 (0.4059)	2.1344 (0.4064)
Heroin	2.7471 (0.5822)	2.9571 (0.522)	2.688 (0.6289)
Cocaine	2.0928 (0.4817)	2.0782 (0.4476)	2.1769 (0.4617)
Magic Mushrooms	1.3383 (0.4843)		
Crack			
Methadone			
Ecstasy			
Amphetamines			
LSD			
Poppers			
Ketamine			
Anabolic Steroids	-1.9414 (1.0203)		
Gas			
Other Drugs			
Tranquillisers	1.511 (0.726)	2.1943 (0.6202)	1.9876 (0.8471)
Methadone			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.5812 (0.3353)	-6.4512 (0.3045)	-6.2018 (0.2856)
Cannabis	2.7722 (0.4268)	2.7288 (0.3844)	2.6675 (0.3697)
Heroin	1.6089 (0.6077)	2.117 (0.5686)	2.0742 (0.5162)
Cocaine			
Magic Mushrooms			
Crack			
Methadone			
Ecstasy	1.6518 (0.4626)	1.0817 (0.4527)	1.2636 (0.4461)
Amphetamines	1.9609 (0.4595)	1.9326 (0.424)	1.9547 (0.4057)
LSD			
Poppers			
Ketamine			
Anabolic Steroids			
Gas			
Other Drugs			
Tranquillisers		1.264 (0.6059)	

Table B.3.3: Table of Estimates of Univariate Logistic Regression Final Models within Drug-trying Response Variables (Table 3)

Ecstasy			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-7.0491 (0.3895)	-6.6318 (0.3117)	-6.372 (0.3093)
Cannabis	2.9169 (0.4501)	2.6454 (0.378)	2.4153 (0.3606)
Heroin			
Cocaine	2.5103 (0.3882)	2.1688 (0.3646)	2.161 (0.3879)
Magic Mushrooms	0.9971 (0.4252)	1.054 (0.3981)	1.063 (0.4036)
Crack			
Methadone	1.4226 (0.5555)		
Ecstasy			
Amphetamines	1.2128 (0.4688)	1.5387 (0.4134)	1.4918 (0.439)
LSD	2.2862 (0.5259)	2.3854 (0.4726)	2.4016 (0.5146)
Poppers			
Ketamine	1.7167 (0.653)	1.7535 (0.5836)	1.6386 (0.5163)
Anabolic Steroids			
Gas	1.2246 (0.3544)	1.1697 (0.3362)	1.0951 (0.3346)
Other Drugs			
Tranquillisers			
Amphetamine			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.4145 (0.3043)	-6.2912 (0.2826)	-6.0457 (0.2674)
Cannabis	2.3824 (0.4165)	2.3813 (0.3836)	2.1558 (0.3596)
Heroin			
Cocaine	1.1801 (0.4338)	1.1134 (0.4103)	1.1136 (0.3959)
Magic Mushrooms	1.6172 (0.3912)	1.583 (0.3788)	1.5306 (0.3939)
Crack			
Methadone	1.7816 (0.4862)	1.6973 (0.4543)	1.6726 (0.4403)
Ecstasy	0.9449 (0.4486)	1.2555 (0.4077)	1.2797 (0.4064)
Amphetamines			
LSD			
Poppers	1.0854 (0.3828)	0.8457 (0.3701)	0.8032 (0.3722)
Ketamine			
Anabolic Steroids			
Gas			
Other Drugs			
Tranquillisers			
LSD			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-7.2787 (0.4565)	-7.0046 (0.3905)	-6.6979 (0.3603)
Cannabis	2.5932 (0.5785)	2.1616 (0.5309)	1.9202 (0.514)
Heroin	1.8542 (0.6248)	2.0859 (0.6072)	1.9769 (0.7262)
Cocaine			
Magic Mushrooms	1.9545 (0.4442)	1.9568 (0.4416)	1.8625 (0.4522)
Crack			
Methadone			
Ecstasy	2.3914 (0.4538)	2.4816 (0.4449)	2.5401 (0.4748)
Amphetamines			
LSD			
Poppers	1.3871 (0.4433)	1.2895 (0.4468)	1.2544 (0.4251)
Ketamine	-2.1615 (0.8623)	-2.0542 (0.8108)	-1.772 (0.7841)
Anabolic Steroids			
Gas			
Other Drugs			
Tranquillisers			

Table B.3.4: Table of Estimates of Univariate Logistic Regression Final Models within Drug-trying Response Variables (Table 4)

Poppers			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-5.349 (0.1772)	-5.3932 (0.1768)	-5.309 (0.1798)
Cannabis	3.1558 (0.2163)	3.1617 (0.2124)	3.1767 (0.208)
Heroin			
Cocaine	1.5417 (0.3074)	1.4797 (0.3058)	1.5445 (0.3028)
Magic Mushrooms	0.68 (0.3155)	0.6333 (0.3058)	
Crack			
Methadone			
Ecstasy			
Amphetamines	1.0371 (0.3578)	0.9434 (0.3417)	0.9299 (0.3861)
LSD	0.9255 (0.4375)	0.8981 (0.4263)	1.0284 (0.4081)
Poppers			
Ketamine			
Anabolic Steroids			
Gas	0.9871 (0.2175)	1.0067 (0.21)	1.0087 (0.2136)
Other Drugs		0.9467 (0.4784)	1.0348 (0.4936)
Tranquillisers			
Ketamine			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.8171 (0.3786)	-6.7563 (0.354)	-6.5628 (0.3429)
Cannabis	3.0986 (0.46)	2.8608 (0.4471)	2.8376 (0.4323)
Heroin		2.114 (0.6027)	2.0136 (0.6124)
Cocaine			
Magic Mushrooms			
Crack			
Methadone			
Ecstasy		1.4392 (0.5284)	1.3335 (0.4935)
Amphetamines	1.814 (0.4737)	1.4766 (0.4927)	1.6181 (0.5255)
LSD		-1.8783 (0.824)	-1.5951 (0.7317)
Poppers			
Ketamine			
Anabolic Steroids			
Gas			
Other Drugs	1.637 (0.6211)	1.5218 (0.67)	
Tranquillisers	2.1775 (0.6089)	1.734 (0.6776)	2.2227 (0.6153)
Anabolic Steroids			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.8057 (0.3482)	-6.7371 (0.3276)	-6.4052 (0.3028)
Cannabis	2.053 (0.4476)	1.7098 (0.4316)	1.8518 (0.4148)
Heroin	1.8887 (0.5867)	1.3435 (0.6578)	1.9483 (0.6655)
Cocaine			
Magic Mushrooms			
Crack			
Methadone			
Ecstasy			
Amphetamines	1.2496 (0.537)	1.1884 (0.5448)	
LSD		1.4418 (0.6513)	1.3857 (0.5845)
Poppers			
Ketamine			
Anabolic Steroids			
Gas	1.8589 (0.4106)	1.9801 (0.3994)	1.681 (0.3731)
Other Drugs	1.697 (0.6703)	1.8425 (0.6425)	1.9379 (0.6956)
Tranquillisers			

Table B.3.5: Table of Estimates of Univariate Logistic Regression Final Models within Drug-trying Response Variables (Table 5)

Gas			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-2.634 (0.0507)	-2.6275 (0.0494)	-2.6213 (0.0494)
Cannabis	0.8094 (0.136)	0.7635 (0.1319)	0.7705 (0.1315)
Heroin	1.7214 (0.4693)	1.0928 (0.4221)	0.9839 (0.3858)
Cocaine	-0.7606 (0.3747)		
Magic Mushrooms	0.8989 (0.2616)	0.8747 (0.255)	0.9361 (0.2566)
Crack			
Methadone			
Ecstasy	0.9273 (0.3357)	0.6495 (0.302)	0.6683 (0.2855)
Amphetamines			
LSD			
Poppers	0.9337 (0.2224)	0.8467 (0.2112)	0.8132 (0.2115)
Ketamine			
Anabolic Steroids	1.7056 (0.4293)	1.7923 (0.4067)	1.49 (0.3794)
Gas			
Other Drugs			
Tranquillisers			
Other Drugs			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.6843 (0.3541)	-6.6148 (0.3338)	-6.2266 (0.3419)
Cannabis	2.3636 (0.4935)	2.6111 (0.4461)	2.0142 (0.498)
Heroin	-2.061 (1.1018)		
Cocaine	1.6333 (0.5405)	1.8947 (0.4736)	1.642 (0.5249)
Magic Mushrooms	1.3862 (0.5324)		
Crack			
Methadone			
Ecstasy			
Amphetamines			
LSD			
Poppers			1.0451 (0.4973)
Ketamine	1.922 (0.5963)		1.3898 (0.5577)
Anabolic Steroids	1.653 (0.7591)		
Gas			
Other Drugs			
Tranquillisers		2.0648 (0.5832)	
Tranquillisers			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.6144 (0.337)	-6.4956 (0.3092)	-6.2404 (0.2924)
Cannabis	1.7768 (0.5199)	1.6056 (0.4879)	1.4099 (0.493)
Heroin			
Cocaine			
Magic Mushrooms	1.9219 (0.5566)		1.8991 (0.5185)
Crack	1.6722 (0.6505)	1.8444 (0.631)	2.0391 (0.6164)
Methadone			
Ecstasy			
Amphetamines		1.5089 (0.5784)	
LSD		1.7114 (0.6476)	
Poppers			
Ketamine	1.8592 (0.6546)	1.5875 (0.6883)	1.7976 (0.6406)
Anabolic Steroids			
Gas			
Other Drugs	1.6866 (0.6835)	1.9432 (0.7147)	1.8307 (0.7085)
Tranquillisers			

B.3.2 Within Response Variables in Saturated Model

Table B.3.6: Table of Estimates of Univariate Logistic Regression Saturated Models within Drug-trying Response Variables (Table 1)

Cannabis			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-2.7948 (0.0544)	-2.7512 (0.0522)	-2.7538 (0.0526)
Cannabis			
Heroin	-0.2742 (0.741)	-0.9856 (0.7907)	-0.8696 (0.795)
Cocaine	1.9528 (0.44)	1.9913 (0.4352)	1.7912 (0.4032)
Magic Mushrooms	1.7251 (0.3155)	1.7608 (0.3071)	1.7126 (0.3004)
Crack	1.3349 (0.5607)	1.3965 (0.5241)	1.3166 (0.5541)
Methadone	2.2087 (0.513)	1.9756 (0.4919)	1.8659 (0.5057)
Ecstasy	2.1388 (0.5099)	1.7925 (0.4585)	1.5767 (0.4282)
Amphetamines	1.1483 (0.5104)	1.33 (0.4675)	1.1095 (0.4367)
LSD	1.9239 (0.6822)	1.08 (0.6419)	0.8744 (0.621)
Poppers	2.9544 (0.2274)	2.9746 (0.2233)	2.9597 (0.2173)
Ketamine	3.0163 (0.4919)	2.8789 (0.4757)	2.7467 (0.4587)
Anabolic Steroids	1.6512 (0.5254)	1.5197 (0.498)	1.4334 (0.5178)
Gas	0.694 (0.1441)	0.6873 (0.1385)	0.7037 (0.1389)
Other Drugs	1.4797 (0.6363)	1.4148 (0.6156)	0.9269 (0.7024)
Tranquillisers	0.4452 (0.6753)	-0.2809 (0.7098)	-0.6016 (0.8085)
Heroin			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-7.0439 (0.3923)	-6.8069 (0.3458)	-6.4266 (0.3378)
Cannabis	0.7079 (0.5992)	0.3732 (0.6045)	0.3997 (0.6187)
Heroin			
Cocaine	2.9367 (0.6374)	2.6848 (0.6504)	2.2743 (0.6988)
Magic Mushrooms	0.9827 (0.6538)	1.0714 (0.6579)	0.9691 (0.684)
Crack	2.5676 (0.6286)	2.8023 (0.6043)	2.5861 (0.6437)
Methadone	0.6438 (0.7563)	1.1038 (0.7559)	1.0027 (0.784)
Ecstasy	-1.0123 (0.7965)	-0.9098 (0.7983)	-0.4558 (0.8036)
Amphetamines	0.1325 (0.7724)	-0.2087 (0.7951)	-0.2657 (0.7675)
LSD	1.3338 (0.8339)	1.8883 (0.7826)	1.4072 (0.8153)
Poppers	0.0509 (0.6867)	0.0595 (0.6897)	0.1772 (0.6423)
Ketamine	0.7769 (1.3284)	2.4662 (0.8922)	2.0979 (1.1125)
Anabolic Steroids	1.8646 (0.8092)	1.8396 (0.7763)	1.9969 (0.8006)
Gas	1.9614 (0.4743)	1.715 (0.4525)	1.3969 (0.4475)
Other Drugs	-2.0068 (1.3497)	-2.4703 (1.4856)	-2.4866 (1.4565)
Tranquillisers	-1.4783 (1.1838)	-2.0494 (1.2095)	-0.7087 (1.4123)
Cocaine			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.274 (0.2838)	-6.181 (0.2669)	-6.0034 (0.2664)
Cannabis	2.505 (0.3694)	2.4145 (0.3602)	2.2683 (0.349)
Heroin	2.9781 (0.6106)	2.751 (0.5928)	2.3667 (0.6003)
Cocaine			
Magic Mushrooms	0.5886 (0.4512)	0.7653 (0.4162)	0.6846 (0.4168)
Crack	2.2425 (0.5815)	2.0233 (0.5325)	2.059 (0.559)
Methadone	-0.214 (0.6329)	-0.273 (0.5829)	-0.0569 (0.5665)
Ecstasy	2.2752 (0.4509)	2.1005 (0.4237)	2.0452 (0.4271)
Amphetamines	1.4347 (0.4669)	1.3409 (0.4335)	1.3095 (0.4296)
LSD	-0.1394 (0.6361)	-0.0678 (0.6017)	0.0496 (0.576)
Poppers	1.5309 (0.3567)	1.4298 (0.3465)	1.3894 (0.3754)
Ketamine	-1.2586 (0.8643)	-1.4818 (0.7762)	-1.2707 (0.7255)
Anabolic Steroids	0.4301 (0.8445)	0.5867 (0.7697)	0.412 (0.7783)
Gas	-0.5286 (0.4033)	-0.3445 (0.3722)	-0.2939 (0.3667)
Other Drugs	1.5425 (0.6189)	1.6959 (0.5811)	1.589 (0.6441)
Tranquillisers	0.3043 (0.7862)	0.3144 (0.7004)	0.2978 (0.8158)

Table B.3.7: Table of Estimates of Univariate Logistic Regression Saturated Models within Drug-trying Response Variables (Table 2)

Magic Mushrooms			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-5.5493 (0.1956)	-5.5807 (0.1957)	-5.4783 (0.1893)
Cannabis	2.2358 (0.278)	2.2627 (0.2712)	2.1918 (0.2676)
Heroin	1.2254 (0.6353)	1.3203 (0.6247)	1.1891 (0.6078)
Cocaine	0.202 (0.4581)	0.367 (0.4334)	0.3503 (0.4302)
Magic Mushrooms			
Crack	1.0317 (0.5755)	0.6152 (0.5718)	0.523 (0.5633)
Methadone	0.2224 (0.5598)	-0.1571 (0.537)	-0.1318 (0.5586)
Ecstasy	0.506 (0.466)	0.7342 (0.4258)	0.774 (0.4221)
Amphetamines	1.4649 (0.4179)	1.4776 (0.3954)	1.3791 (0.4062)
LSD	1.4529 (0.5195)	1.5431 (0.4977)	1.3794 (0.5432)
Poppers	0.459 (0.3419)	0.4114 (0.3305)	0.3661 (0.3291)
Ketamine	0.2292 (0.6693)	0.462 (0.5759)	0.3459 (0.6251)
Anabolic Steroids	0.1145 (0.7093)	-0.149 (0.7067)	-0.0423 (0.7743)
Gas	1.055 (0.2736)	1.1162 (0.2658)	1.1294 (0.2554)
Other Drugs	1.2269 (0.5764)	0.8716 (0.5873)	0.7376 (0.6536)
Tranquillisers	0.9762 (0.7035)	0.6385 (0.666)	1.0104 (0.6944)
Crack			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.7001 (0.3443)	-6.5114 (0.3201)	-6.3966 (0.308)
Cannabis	2.0624 (0.4781)	2.0025 (0.4448)	1.9116 (0.4328)
Heroin	2.6703 (0.6059)	2.9255 (0.5647)	2.7354 (0.5882)
Cocaine	2.0013 (0.5555)	1.8936 (0.5162)	1.9724 (0.5322)
Magic Mushrooms	1.0469 (0.557)	0.7139 (0.5667)	0.6806 (0.5479)
Crack			
Methadone	0.1843 (0.7007)	0.9009 (0.624)	0.7979 (0.6364)
Ecstasy	-0.0604 (0.7174)	-0.5113 (0.6987)	-0.7054 (0.73)
Amphetamines	0.1177 (0.6865)	0.0311 (0.6634)	0.1143 (0.7357)
LSD	0.0857 (0.771)	0.3048 (0.8063)	0.3864 (0.794)
Poppers	6e-04 (0.5582)	0.1412 (0.5266)	0.0514 (0.5582)
Ketamine	-0.4973 (0.9531)	-0.1293 (0.8455)	-0.0668 (1.0517)
Anabolic Steroids	-2.0951 (1.0686)	-1.9623 (1.0166)	-1.7165 (1.233)
Gas	0.6927 (0.4663)	0.5352 (0.4693)	0.6429 (0.4552)
Other Drugs	1.0734 (0.7724)	0.9984 (0.7605)	1.186 (0.8158)
Tranquillisers	1.4205 (0.7829)	1.8793 (0.6978)	1.6295 (0.8634)
Methadone			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.6585 (0.3412)	-6.4932 (0.3073)	-6.2618 (0.2874)
Cannabis	2.5313 (0.4505)	2.4448 (0.414)	2.3345 (0.4147)
Heroin	1.2172 (0.7502)	1.2414 (0.7926)	1.1821 (0.8311)
Cocaine	0.2118 (0.5834)	0.2182 (0.5606)	0.2803 (0.5473)
Magic Mushrooms	0.8293 (0.5038)	0.3696 (0.5042)	0.2841 (0.5331)
Crack	-0.0145 (0.7294)	0.8164 (0.6631)	0.7571 (0.6467)
Methadone			
Ecstasy	1.3814 (0.5577)	0.8531 (0.5503)	0.8308 (0.5392)
Amphetamines	1.7542 (0.4734)	1.6518 (0.4509)	1.5749 (0.4396)
LSD	-0.9183 (0.7498)	-0.5784 (0.7241)	-0.6541 (0.6791)
Poppers	0.2006 (0.4731)	0.3428 (0.4442)	0.3033 (0.4317)
Ketamine	0.3231 (0.7249)	0.1628 (0.6815)	0.2881 (0.6116)
Anabolic Steroids	0.8408 (0.7444)	1.1036 (0.7085)	1.0377 (0.7453)
Gas	0.5814 (0.4041)	0.5288 (0.3897)	0.5214 (0.4141)
Other Drugs	-0.2712 (0.8759)	0.3239 (0.7433)	0.5823 (0.8225)
Tranquillisers	0.0904 (0.8018)	0.7311 (0.707)	0.6351 (0.8455)

Table B.3.8: Table of Estimates of Univariate Logistic Regression Saturated Models within Drug-trying Response Variables (Table 3)

Ecstasy			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-7.0424 (0.3873)	-6.6289 (0.3124)	-6.3954 (0.3054)
Cannabis	2.8287 (0.4613)	2.5811 (0.4032)	2.3683 (0.378)
Heroin	-0.5187 (0.8156)	-0.382 (0.8415)	-0.0508 (0.8609)
Cocaine	2.426 (0.4299)	2.2037 (0.4165)	2.1662 (0.4271)
Magic Mushrooms	0.9322 (0.4433)	0.9857 (0.4179)	0.9961 (0.4197)
Crack	0.4857 (0.6791)	-0.2336 (0.6866)	-0.4804 (0.6785)
Methadone	1.5059 (0.5735)	0.9378 (0.5794)	0.9078 (0.594)
Ecstasy			
Amphetamines	1.1483 (0.482)	1.4115 (0.4388)	1.3838 (0.4643)
LSD	2.2252 (0.5415)	2.4588 (0.5216)	2.4616 (0.5506)
Poppers	0.2232 (0.4162)	0.1114 (0.3954)	0.1502 (0.3994)
Ketamine	1.5947 (0.7237)	1.6218 (0.6519)	1.4347 (0.5733)
Anabolic Steroids	-0.4639 (0.8602)	-0.8987 (0.9345)	-1.0751 (0.9616)
Gas	1.28 (0.3627)	1.1995 (0.345)	1.1255 (0.3458)
Other Drugs	0.2973 (0.7554)	0.0089 (0.7325)	0.1865 (0.7252)
Tranquillisers	0.595 (0.8608)	0.6691 (0.8198)	0.7385 (0.9081)
Amphetamine			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.4595 (0.3089)	-6.3023 (0.2855)	-6.0663 (0.2634)
Cannabis	2.3087 (0.4233)	2.3715 (0.3863)	2.1052 (0.3672)
Heroin	0.0693 (0.7807)	0.1368 (0.7712)	-0.0829 (0.7515)
Cocaine	1.2797 (0.4655)	1.1007 (0.4508)	1.0959 (0.4486)
Magic Mushrooms	1.5984 (0.4247)	1.4932 (0.4126)	1.4089 (0.4286)
Crack	-0.2674 (0.7333)	-0.3433 (0.6671)	-0.1565 (0.7516)
Methadone	1.7028 (0.5149)	1.4939 (0.4885)	1.4595 (0.4826)
Ecstasy	0.7967 (0.4992)	1.1805 (0.4567)	1.1604 (0.4834)
Amphetamines			
LSD	-0.1329 (0.6737)	-0.2799 (0.661)	-0.0801 (0.6539)
Poppers	1.0667 (0.3984)	0.7124 (0.3975)	0.6765 (0.4175)
Ketamine	1.2131 (0.7222)	0.8987 (0.6615)	1.1133 (0.7125)
Anabolic Steroids	0.1623 (0.7751)	0.6676 (0.78)	0.8102 (0.7748)
Gas	0.4141 (0.384)	0.1791 (0.3797)	0.1859 (0.3904)
Other Drugs	-0.1087 (0.7409)	-0.2168 (0.7186)	-0.0624 (0.7033)
Tranquillisers	-0.4614 (0.7971)	0.5072 (0.7753)	0.1879 (0.8047)
LSD			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-7.281 (0.4589)	-7.0036 (0.3936)	-6.7234 (0.3649)
Cannabis	2.5198 (0.5885)	2.0614 (0.5478)	1.8113 (0.5256)
Heroin	1.5662 (0.7617)	1.9424 (0.7651)	1.7723 (0.7777)
Cocaine	0.4469 (0.5402)	0.3558 (0.5499)	0.4078 (0.5229)
Magic Mushrooms	1.8348 (0.467)	1.8741 (0.4737)	1.7722 (0.4943)
Crack	0.3658 (0.7596)	0.2843 (0.7664)	0.3029 (0.727)
Methadone	-0.338 (0.6941)	-0.2498 (0.6721)	-0.289 (0.6151)
Ecstasy	2.1352 (0.5184)	2.3686 (0.5213)	2.4163 (0.5445)
Amphetamines	0.2453 (0.6065)	0.0851 (0.5983)	0.1311 (0.5872)
LSD			
Poppers	1.1914 (0.4688)	1.0481 (0.4851)	0.9956 (0.4771)
Ketamine	-2.7831 (1.0108)	-2.8887 (0.9047)	-2.5525 (0.9153)
Anabolic Steroids	0.8553 (0.8454)	1.336 (0.8821)	1.2999 (0.9197)
Gas	0.0631 (0.4807)	-0.2206 (0.4947)	-0.2377 (0.5006)
Other Drugs	0.7571 (0.7663)	1.1293 (0.7251)	1.1309 (0.7979)
Tranquillisers	0.5225 (0.7777)	0.8841 (0.775)	0.8188 (0.7249)

Table B.3.9: Table of Estimates of Univariate Logistic Regression Saturated Models within Drug-trying Response Variables (Table 4)

Poppers			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-5.3525 (0.1775)	-5.3977 (0.1775)	-5.3126 (0.1793)
Cannabis	3.1355 (0.2183)	3.1586 (0.214)	3.1405 (0.2114)
Heroin	-0.0565 (0.6117)	0.1366 (0.5783)	0.2787 (0.5649)
Cocaine	1.5416 (0.355)	1.4157 (0.3448)	1.3699 (0.3653)
Magic Mushrooms	0.6192 (0.3309)	0.5345 (0.3205)	0.4696 (0.3202)
Crack	-0.3272 (0.5533)	0.0159 (0.5016)	-0.0855 (0.5522)
Methadone	0.128 (0.5028)	0.4149 (0.454)	0.3596 (0.4227)
Ecstasy	-0.1759 (0.4199)	-0.1259 (0.3924)	-0.0736 (0.4137)
Amphetamines	0.9807 (0.383)	0.7909 (0.369)	0.7126 (0.3996)
LSD	0.9343 (0.477)	0.8621 (0.4736)	0.7845 (0.4616)
Poppers			
Ketamine	0.064 (0.5821)	0.3239 (0.5281)	0.1499 (0.5297)
Anabolic Steroids	0.0309 (0.6006)	-0.1242 (0.6051)	-0.1033 (0.5524)
Gas	1.0065 (0.2223)	1.01 (0.2149)	0.9496 (0.2162)
Other Drugs	0.8301 (0.5359)	0.7514 (0.5128)	0.8521 (0.5252)
Tranquillisers	0.6136 (0.6555)	0.7108 (0.6155)	0.7118 (0.671)
Ketamine			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.8372 (0.3814)	-6.7553 (0.3567)	-6.5873 (0.3507)
Cannabis	2.8991 (0.484)	2.7851 (0.4584)	2.7238 (0.4461)
Heroin	1.5407 (0.9374)	2.4987 (0.7563)	2.1727 (0.8975)
Cocaine	-0.56 (0.7205)	-0.9446 (0.6949)	-0.7125 (0.635)
Magic Mushrooms	0.6278 (0.5823)	0.6741 (0.5321)	0.5803 (0.5925)
Crack	0.2467 (0.8313)	0.2674 (0.766)	0.2602 (0.8316)
Methadone	0.4479 (0.6862)	0.5599 (0.5943)	0.6077 (0.545)
Ecstasy	1.1951 (0.6464)	1.5146 (0.6018)	1.3076 (0.5472)
Amphetamines	1.4595 (0.5827)	1.4067 (0.5416)	1.4898 (0.5723)
LSD	-1.885 (0.9464)	-1.9168 (0.8498)	-1.7559 (0.7807)
Poppers	-0.0012 (0.5492)	0.1694 (0.5152)	0.0518 (0.496)
Ketamine			
Anabolic Steroids	-0.3173 (1.029)	-0.8381 (0.9512)	-0.6154 (1.3286)
Gas	0.052 (0.4753)	-0.2095 (0.4703)	-0.132 (0.4904)
Other Drugs	1.6407 (0.7152)	1.7751 (0.6885)	1.5914 (0.7287)
Tranquillisers	1.8483 (0.7255)	1.5951 (0.7063)	1.7398 (0.6738)
Anabolic Steroids			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.8545 (0.3588)	-6.8493 (0.3472)	-6.504 (0.3168)
Cannabis	1.9626 (0.4787)	1.8369 (0.451)	1.7808 (0.4429)
Heroin	1.8799 (0.8234)	1.77 (0.8432)	2.01 (0.8081)
Cocaine	0.2675 (0.7589)	0.4316 (0.7256)	0.3483 (0.7575)
Magic Mushrooms	0.2814 (0.6655)	-0.0049 (0.6577)	0.1127 (0.7049)
Crack	-1.6549 (1.1061)	-1.2606 (1.0178)	-1.2857 (1.1667)
Methadone	0.516 (0.723)	0.6573 (0.6756)	0.7534 (0.665)
Ecstasy	-0.5789 (0.861)	-1.0174 (0.8498)	-1.0718 (1.0617)
Amphetamines	0.9022 (0.6985)	1.0588 (0.668)	1.0071 (0.6967)
LSD	1.1668 (0.819)	1.6142 (0.8425)	1.4716 (0.7887)
Poppers	-0.0527 (0.59)	-0.2745 (0.5997)	-0.2035 (0.6007)
Ketamine	-0.3334 (1.0241)	-0.4494 (0.9642)	-0.3434 (1.2313)
Anabolic Steroids			
Gas	1.9121 (0.4259)	2.1144 (0.4101)	1.7772 (0.3846)
Other Drugs	1.5994 (0.7265)	1.7678 (0.6662)	1.7351 (0.7889)
Tranquillisers	1.3639 (0.8165)	1.6723 (0.7601)	1.2483 (0.8262)

Table B.3.10: Table of Estimates of Univariate Logistic Regression Saturated Models within Drug-trying Response Variables (Table 5)

Gas			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-2.6366 (0.0507)	-2.6317 (0.0494)	-2.6249 (0.0495)
Cannabis	0.8123 (0.1364)	0.7833 (0.1323)	0.7812 (0.133)
Heroin	1.7007 (0.4825)	1.4627 (0.4633)	1.1644 (0.4259)
Cocaine	-0.8527 (0.3906)	-0.5946 (0.3595)	-0.4711 (0.3502)
Magic Mushrooms	0.9327 (0.2702)	1.0158 (0.2623)	1.0309 (0.2557)
Crack	0.3949 (0.4713)	0.1878 (0.4719)	0.3523 (0.4679)
Methadone	0.2282 (0.4311)	0.1315 (0.4124)	0.1982 (0.4349)
Ecstasy	1.0052 (0.3617)	1.0404 (0.3372)	0.9467 (0.3316)
Amphetamines	0.1805 (0.3779)	0.0018 (0.3658)	0.0581 (0.3684)
LSD	-0.7233 (0.496)	-0.781 (0.4743)	-0.6307 (0.458)
Poppers	0.9518 (0.2256)	0.9567 (0.2169)	0.8876 (0.2167)
Ketamine	-0.5633 (0.5371)	-0.8208 (0.5035)	-0.5875 (0.4936)
Anabolic Steroids	1.7365 (0.4333)	1.8901 (0.4134)	1.5783 (0.3906)
Gas			
Other Drugs	-0.0266 (0.5117)	0.0953 (0.4828)	0.0288 (0.4981)
Tranquillisers	0.3684 (0.5387)	0.1316 (0.5268)	0.0412 (0.502)
Other Drugs			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.7083 (0.3574)	-6.6539 (0.342)	-6.2806 (0.3505)
Cannabis	2.0803 (0.533)	2.0967 (0.4979)	1.7552 (0.539)
Heroin	-3.348 (1.3762)	-3.5481 (1.6996)	-3.0873 (1.525)
Cocaine	1.0463 (0.6254)	1.3033 (0.6072)	1.129 (0.6532)
Magic Mushrooms	1.142 (0.5776)	0.7399 (0.6114)	0.6086 (0.6548)
Crack	1.0911 (0.8122)	1.1775 (0.8858)	1.0312 (0.8388)
Methadone	-0.9012 (0.8818)	-0.3984 (0.7958)	0.0445 (0.7676)
Ecstasy	0.0983 (0.7354)	-0.2364 (0.7243)	0.0708 (0.725)
Amphetamines	0.4069 (0.6613)	0.2652 (0.6481)	0.1716 (0.6643)
LSD	0.1473 (0.773)	0.4953 (0.7552)	0.545 (0.8355)
Poppers	0.9673 (0.5617)	0.7966 (0.5629)	0.8479 (0.5583)
Ketamine	1.629 (0.7365)	1.4771 (0.7504)	1.2736 (0.8242)
Anabolic Steroids	1.8255 (0.8021)	1.8422 (0.7843)	1.7604 (0.8909)
Gas	0.1357 (0.5194)	0.1843 (0.506)	0.079 (0.5311)
Other Drugs			
Tranquillisers	0.734 (0.772)	1.0224 (0.7761)	0.8927 (0.8881)
Tranquillisers			
	Complete Case	MICE, Scheme 1	MICE, Scheme 2
(Intercept)	-6.689 (0.3479)	-6.5551 (0.3173)	-6.2724 (0.299)
Cannabis	1.1917 (0.6103)	0.9601 (0.5832)	0.7793 (0.5805)
Heroin	-1.9148 (1.2312)	-2.2281 (1.1439)	-0.7951 (1.4433)
Cocaine	0.5486 (0.7773)	0.1061 (0.7739)	0.0946 (0.9906)
Magic Mushrooms	1.2438 (0.6563)	0.804 (0.6838)	1.0792 (0.754)
Crack	1.4657 (0.8057)	2.056 (0.7152)	1.6482 (0.8318)
Methadone	-0.4803 (0.8894)	0.435 (0.7384)	0.412 (0.8257)
Ecstasy	0.5553 (0.8233)	0.7707 (0.7972)	0.8861 (0.9086)
Amphetamines	0.6089 (0.7318)	0.9048 (0.6862)	0.4311 (0.808)
LSD	0.7902 (0.8257)	1.1165 (0.8168)	0.8254 (0.802)
Poppers	0.5692 (0.6532)	0.5474 (0.6394)	0.6049 (0.7034)
Ketamine	1.7461 (0.8025)	1.4993 (0.8335)	1.5892 (0.8314)
Anabolic Steroids	1.3087 (0.886)	1.5637 (0.8202)	1.0878 (1.0254)
Gas	0.8159 (0.5262)	0.5245 (0.5236)	0.2998 (0.5346)
Other Drugs	1.1385 (0.7828)	1.4342 (0.7685)	1.2766 (0.8606)
Tranquillisers			

B.4 Univariate Logistic Regression with Covariates

Table B.4.1: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 1)

Variable	Levels	Type	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack
(Intercept)	(Intercept)	Linear	-9.1083 (0.7778)	-5.0633 (0.5648)	-6.8191 (0.4485)	-6.3036 (0.3779)	-5.5787 (0.3989)
CgFam1	Linear	Linear					
CgFam1	Against	Factor	—***—	—***—	—***—	—***—	—***—
	Middle	Factor					-0.4667 (0.8292)
	For	Factor					1.6797 (0.5699)
CgStat1	Linear	Linear					
CgStat1	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Tried/Before	Factor	2.3924 (0.2106)		0.8485 (0.5044)	0.8907 (0.4286)	
	Current-light	Factor	0.3877 (0.9333)		0.7048 (0.6719)	1.77 (0.4504)	
	Current-moderate	Factor	0.8510 (0.8950)		0.2004 (0.9238)	1.9453 (0.5062)	
	Current-heavy	Factor	1.5900 (0.9842)		2.4585 (0.7452)	1.2028 (0.5757)	
CgPk1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Packet	Factor					
	Hand-rolled	Factor					
	Both	Factor					
CgGet1	Linear	Linear					
CgGet1	None	Factor	—***—	—***—	—***—	—***—	—***—
	1	Factor					-1.3466 (0.6184)
	>1	Factor					-2.2321 (0.9899)

Table B.4.2: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 2)

Variable	Levels	Type	Methadone	Ecstasy	Amphetamines	LSD	Poppers
(Intercept)	(Intercept)	Linear	-5.9719 (0.6938)	-7.8489 (0.6851)	-6.8023 (0.5839)	-7.4109 (0.9518)	-6.2012 (0.2465)
CgFam1	Linear	Linear					
CgFam1	Against	Factor	—***—	—***—	—***—	—***—	—***—
	Middle	Factor					
	For	Factor					
CgStat1	Linear	Linear					
CgStat1	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Tried/Before	Factor			0.7981 (0.576)	2.9551 (1.0806)	
	Current-light	Factor			1.6061 (0.6103)	1.3678 (1.2515)	
	Current-moderate	Factor			2.4655 (0.6623)	3.0956 (1.1951)	
	Current-missing	Factor			2.3165 (0.7436)	3.0572 (1.1857)	
CgPk1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Packet	Factor					
	Hand-rolled	Factor					
	Both	Factor					
CgGet1	Linear	Linear					
CgGet1	None	Factor	—***—	—***—	—***—	—***—	—***—
	1	Factor					
	>1	Factor					

Table B.4.3: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 3)

Variable	Levels	Type	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
(Intercept)	(Intercept)	Linear	-6.6874 (0.3444)	-6.5295 (0.6114)	-4.863 (0.432)	-6.7733 (0.427)	-5.2069 (0.5536)
CgFam1	Linear	Linear					
CgFam1	Against	Factor	—***—	—***—	—***—	—***—	—***—
	Middle	Factor					
	For	Factor					
CgStat1	Linear	Linear				0.7205 (0.1543)	
CgStat1	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Tried/Before	Factor					
	Current-light	Factor					
	Current-moderate	Factor					
	Current-missing	Factor					
CgPk1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Packet	Factor					
	Hand-rolled	Factor					
	Both	Factor					
CgGet1	Linear	Linear					1.3055 (0.4146)
CgGet1	None	Factor	—***—	—***—	—***—	—***—	—***—
	1	Factor					
	>1	Factor					

Table B.4.4: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 4)

Variable	Levels	Type	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack
CgGet2	Linear	Linear					
CgGet2	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops Only	Factor			-0.0098 (0.5337)		
	1	Factor			-0.8262 (0.5909)		
	>1	Factor			-3.1224 (1.1833)		
CgGet3	Linear	Linear					
CgGet3	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops/People Only	Factor					
	1	Factor					
	>1	Factor					
CgGet	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops	Factor	2.4625 (0.9158)				
	People	Factor	2.8112 (0.9337)				
	Given	Factor	2.7422 (0.9240)				
	Mixture	Factor	3.1510 (0.9147)				
CgPp1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Other relatives only	Factor					
	Friends only	Factor					
	Family members only	Factor					
	Mixture	Factor					
CgWho1	Linear	Linear					
CgWho1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Smoke, outside	Factor	0.3740 (0.1542)				
	Smoke, inside	Factor	0.2764 (0.1617)				
CgBuyF1	Linear	Linear					
CgBuyF1	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor	0.5746 (0.2125)				
	Ocassional	Factor	0.5694 (0.2450)				
	Frequent	Factor	-0.8214 (0.5323)				
CgEstim	Linear	Linear					
CgEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor	1.7097 (0.7087)				
	Half	Factor	1.9892 (0.7124)				
	Most	Factor	1.7180 (0.7158)				
	All	Factor	1.5148 (0.8343)				

Table B.4.5: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 5)

Variable	Levels	Type	Methadone	Ecstasy	Amphetamines	LSD	Poppers
CgGet2	Linear	Linear					
CgGet2	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops Only	Factor					
	1	Factor					
	>1	Factor					
CgGet3	Linear	Linear					
CgGet3	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops/People Only	Factor					
	1	Factor					
	>1	Factor					
CgGet	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops	Factor					
	People	Factor					
	Given	Factor					
	Mixture	Factor					
CgPp1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Other relatives only	Factor					
	Friends only	Factor					
	Family members only	Factor					
	Mixture	Factor					
CgWho1	Linear	Linear					
CgWho1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Smoke, outside	Factor					
	Smoke, inside	Factor					
CgBuyF1	Linear	Linear					
CgBuyF1	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor			1.0748 (0.4378)		1.3167 (0.2713)
	Ocassional	Factor			-0.265 (0.4784)		0.7934 (0.2773)
	Frequent	Factor			-0.3604 (0.7156)		1.4491 (0.4513)
CgEstim	Linear	Linear					
CgEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor	-0.8527 (0.8235)				
	Half	Factor	0.2871 (0.7748)				
	Most	Factor	0.8707 (0.7879)				
	All	Factor	-1.8616 (1.4452)				

Table B.4.6: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 6)

Variable	Levels	Type	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
CgGet2	Linear	Linear					
CgGet2	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops Only	Factor					
	1	Factor					
	>1	Factor					
CgGet3	Linear	Linear					
CgGet3	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops/People Only	Factor					
	1	Factor					
	>1	Factor					
CgGet	None	Factor	—***—	—***—	—***—	—***—	—***—
	Shops	Factor					
	People	Factor					
	Given	Factor					
	Mixture	Factor					
CgPp1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Other relatives only	Factor			-0.1008 (0.2203)		
	Friends only	Factor			0.4565 (0.1971)		
	Family members only	Factor			0.3827 (0.2425)		
	Mixture	Factor			0.561 (0.1888)		
CgWho1	Linear	Linear					
CgWho1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Smoke, outside	Factor		0.9973 (0.4955)	-0.2887 (0.1275)		
	Smoke, inside	Factor		1.1378 (0.4921)	-0.5167 (0.1466)		
CgBuyF1	Linear	Linear					
CgBuyF1	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor	1.4727 (0.5236)		0.1711 (0.2231)		
	Ocassional	Factor	1.7147 (0.4882)		-0.8203 (0.2632)		
	Frequent	Factor	-0.7543 (1.0691)		-0.5832 (0.4595)		
CgEstim	Linear	Linear					
CgEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor		-0.4639 (0.6579)			
	Half	Factor		-0.6872 (0.8203)			
	Most	Factor		-2.2529 (0.9974)			
	All	Factor		0.9998 (1.0017)			

Table B.4.7: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 7)

Variable	Levels	Type	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack
LsAlc	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
AlPar1	Linear	Linear	-0.3216 (0.1502)				
AlPar1	Against	Factor	—***—	—***—	—***—	—***—	—***—
	Middle	Factor					
	For	Factor					
CgPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, relatives	Factor					
	Pros, police	Factor					
	Both	Factor					
CgIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive	Factor					
	Interactive	Factor					
	Both	Factor					
LsSmk	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			0.7548 (0.3567)		
AlLast	Linear	Linear					
AlLast	Never	Factor	—***—	—***—	—***—	—***—	—***—
	up to 1 month ago	Factor		0.9663 (0.6406)			
	4 weeks to 1 week ago	Factor		-0.1754 (1.0554)			
	previous week	Factor		1.742 (0.5366)			
AlFreq2	Linear	Linear				0.2906 (0.0831)	
AlFreq2	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Ex-drinker	Factor	1.2970 (0.3812)				
	Few a year	Factor	0.9372 (0.2811)				
	Once a month	Factor	1.2061 (0.2951)				
	Once a fortnight to once a week	Factor	1.6051 (0.2989)				
	Once a week, 1-2 days previous week	Factor	1.2900 (0.3234)				
	Once a week, >2 days previous week	Factor	1.4360 (0.3844)				
AlBnPub	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
AlEstim	Linear	Linear					
AlEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor					
	Half	Factor					
	Most	Factor					
	All	Factor					

Table B.4.8: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 8)

Variable	Levels	Type	Methadone	Ecstasy	Amphetamines	LSD	Poppers
LsAlc	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	-1.6994 (0.5774)				
AlPar1	Linear	Linear					
AlPar1	Against	Factor	—***—	—***—	—***—	—***—	—***—
	Middle	Factor					
	For	Factor					
CgPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, relatives	Factor					
	Pros, police	Factor					
	Both	Factor					
CgIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive	Factor					
	Interactive	Factor					
	Both	Factor					
LsSmk	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.26 (0.6067)				
AlLast	Linear	Linear					
AlLast	Never	Factor	—***—	—***—	—***—	—***—	—***—
	up to 1 month ago	Factor					
	4 weeks to 1 week ago	Factor					
	previous week	Factor					
AlFreq2	Linear	Linear					0.4115 (0.0645)
AlFreq2	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Ex-drinker	Factor					
	Few a year	Factor					
	Once a month	Factor					
	Once a fortnight to once a week	Factor					
	Once a week, 1-2 days previous week	Factor					
	Once a week, >2 days previous week	Factor					
AlBnPub	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
AlEstim	Linear	Linear					
AlEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor					
	Half	Factor					
	Most	Factor					
	All	Factor					

Table B.4.9: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 9)

Variable	Levels	Type	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
LsAlc	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			0.6697 (0.1716)		
AlPar1	Linear	Linear					
AlPar1	Against	Factor	—***—	—***—	—***—	—***—	—***—
	Middle	Factor			0.21 (0.1206)		
	For	Factor			-0.7205 (0.4403)		
CgPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, relatives	Factor			-0.1853 (0.1275)		
	Pros, police	Factor			-0.2754 (0.6844)		
	Both	Factor			-0.5128 (0.1645)		
CgIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive	Factor			0.4065 (0.1921)		
	Interactive	Factor			0.5144 (0.2593)		
	Both	Factor			0.8213 (0.1634)		
LsSmk	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
AlLast	Linear	Linear					
AlLast	Never	Factor	—***—	—***—	—***—	—***—	—***—
	up to 1 month ago	Factor					
	4 weeks to 1 week ago	Factor					
	previous week	Factor					
AlFreq2	Linear	Linear					
AlFreq2	Never	Factor	—***—	—***—	—***—	—***—	—***—
	Ex-drinker	Factor					
	Few a year	Factor					
	Once a month	Factor					
	Once a fortnight to once a week	Factor					
	Once a week, 1-2 days previous week	Factor					
	Once a week, >2 days previous week	Factor					
AlBnPub	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			0.2807 (0.1022)		0.9311 (0.4422)
AlEstim	Linear	Linear					
AlEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor			-0.055 (0.1909)		
	Half	Factor			-0.5442 (0.2218)		
	Most	Factor			-0.3637 (0.2286)		
	All	Factor			-0.062 (0.2846)		

Table B.4.10: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 10)

Variable	Levels	Type	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack
AlBuy1	Linear	Linear					
AlBuy1	0 sources	Factor	—***—	—***—	—***—	—***—	—***—
	1 sources	Factor					
	2 sources	Factor					
	>= 3 sources	Factor					
AlBuy2	Linear	Linear					
AlBuy2	None	Factor	—***—	—***—	—***—	—***—	—***—
	From shops	Factor					
	1	Factor					
	>1	Factor					
AlBuy	None	Factor	—***—	—***—	—***—	—***—	—***—
	places	Factor					
	family members	Factor					
	both	Factor					
AlUs1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Own	Factor					
	Other people and friends	Factor					
	Family members	Factor					
	Both	Factor					
AlUs2	None	Factor	—***—	—***—	—***—	—***—	—***—
	Pub	Factor					
	Home/party	Factor					
	Other place	Factor					
	Mixture	Factor					
Al4W1	None in last 4 weeks	Factor	—***—	—***—	—***—	—***—	—***—
	Drink, no issue	Factor					
	Drink, health issue	Factor					
	Drink, aggresive and other issue	Factor					
	Drink, both	Factor					
AlWhy1	No reasons	Factor	—***—	—***—	—***—	—***—	—***—
	Feel better	Factor					
	Socialise	Factor					
	Both	Factor					
AlWho1	Linear	Linear					
AlWho1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Smoke, outside	Factor					
	Smoke, inside	Factor					

Table B.4.11: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 11)

Variable	Levels	Type	Methadone	Ecstasy	Amphetamines	LSD	Poppers
AlBuy1	Linear	Linear					
AlBuy1	0 sources	Factor	—***—	—***—	—***—	—***—	—***—
	1 sources	Factor					
	2 sources	Factor					
	>= 3 sources	Factor					
AlBuy2	Linear	Linear					
AlBuy2	None	Factor	—***—	—***—	—***—	—***—	—***—
	From shops	Factor			1.5661 (0.4852)		
	1	Factor			0.5852 (0.5044)		
	>1	Factor			0.9145 (0.5561)		
AlBuy	None	Factor	—***—	—***—	—***—	—***—	—***—
	places	Factor					
	family members	Factor					
	both	Factor					
AlUs1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Own	Factor					
	Other people and friends	Factor					
	Family members	Factor					
	Both	Factor					
AlUs2	None	Factor	—***—	—***—	—***—	—***—	—***—
	Pub	Factor					
	Home/party	Factor					
	Other place	Factor					
	Mixture	Factor					
Al4W1	None in last 4 weeks	Factor	—***—	—***—	—***—	—***—	—***—
	Drink, no issue	Factor		1.4003 (0.5856)			
	Drink, health issue	Factor		1.1165 (0.7178)			
	Drink, aggresive and other issue	Factor		1.6093 (0.5775)			
	Drink, both	Factor		2.1923 (0.5224)			
AlWhy1	No reasons	Factor	—***—	—***—	—***—	—***—	—***—
	Feel better	Factor					
	Socialise	Factor					
	Both	Factor					
AlWho1	Linear	Linear					
AlWho1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Smoke, outside	Factor					
	Smoke, inside	Factor					

Table B.4.12: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 12)

Variable	Levels	Type	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
AlBuy1	Linear	Linear					-1.1832 (0.4313)
AlBuy1	0 sources	Factor	—***—	—***—	—***—	—***—	—***—
	1 sources	Factor					
	2 sources	Factor					
	>= 3 sources	Factor					
AlBuy2	Linear	Linear					
AlBuy2	None	Factor	—***—	—***—	—***—	—***—	—***—
	From shops	Factor					
	1	Factor					
	>1	Factor					
AlBuy	None	Factor	—***—	—***—	—***—	—***—	—***—
	places	Factor					
	family members	Factor					
	both	Factor					
AlUs1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Own	Factor					
	Other people and friends	Factor					
	Family members	Factor					
	Both	Factor					
AlUs2	None	Factor	—***—	—***—	—***—	—***—	—***—
	Pub	Factor			-0.1682 (0.5782)		
	Home/party	Factor			0.2793 (0.1677)		
	Other place	Factor			0.6213 (0.1822)		
	Mixture	Factor			0.5032 (0.1721)		
Al4W1	None in last 4 weeks	Factor	—***—	—***—	—***—	—***—	—***—
	Drink, no issue	Factor	0.6659 (0.1978)				
	Drink, health issue	Factor	0.7319 (0.2585)				
	Drink, aggresive and other issue	Factor	0.5437 (0.2685)				
	Drink, both	Factor	0.5083 (0.2297)				
AlWhy1	No reasons	Factor	—***—	—***—	—***—	—***—	—***—
	Feel better	Factor					
	Socialise	Factor					
	Both	Factor					
AlWho1	Linear	Linear					
AlWho1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Smoke, outside	Factor			-0.326 (0.1943)		
	Smoke, inside	Factor			0.1128 (0.1457)		

Table B.4.13: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 13)

Variable	Levels	Type	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack
AlPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, other relatives	Factor					
	Pros, police	Factor					
	Both	Factor					
AlIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive Media	Factor					
	Interactive Media	Factor					
	Both	Factor					
DgPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, other relatives	Factor	0.4807 (0.1568)				
	Pros, police	Factor	-0.3073 (0.7433)				
	Both	Factor	0.0875 (0.1847)				
DgIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive Media	Factor					-2.2219 (1.0415)
	Interactive Media	Factor					0.4146 (0.5789)
	Both	Factor					-1.2256 (0.5553)
DgEstim	Linear	Linear				0.3509 (0.1342)	
DgEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor	0.6854 (0.1876)				
	Half	Factor	1.2084 (0.2438)				
	Most/All	Factor	2.0144 (0.3213)				
Books1	Linear	Linear		-1.316 (0.2958)			
Books1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor					
	Lots	Factor					
LsDrg	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
age	age	Linear	0.3752 (0.0717)			-0.2364 (0.1129)	
gender	Boy	Factor	—***—	—***—	—***—	—***—	—***—
	Girl	Factor	-0.9074 (0.1336)				
FSM1	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
TruantN	Linear	Linear					
TruantN	No	Factor	—***—	—***—	—***—	—***—	—***—
	Played truant a year ago	Factor	0.5708 (0.2324)				
	1/2	Factor	0.112 (0.1817)				
	>=3	Factor	0.3093 (0.2230)				

Table B.4.14: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 14)

Variable	Levels	Type	Methadone	Ecstasy	Amphetamines	LSD	Poppers
AlPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, other relatives	Factor					
	Pros, police	Factor					
	Both	Factor					
AlIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive Media	Factor					
	Interactive Media	Factor					
	Both	Factor					
DgPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, other relatives	Factor					
	Pros, police	Factor					
	Both	Factor					
DgIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive Media	Factor					
	Interactive Media	Factor					
	Both	Factor					
DgEstim	Linear	Linear					
DgEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor			0.6409 (0.5551)		
	Half	Factor			1.5107 (0.5865)		
	Most/All	Factor			0.1926 (0.7356)		
Books1	Linear	Linear					
Books1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor					
	Lots	Factor					
LsDrg	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			-0.7722 (0.3502)		
age	age	Linear				-0.4102 (0.1867)	
gender	Boy	Factor	—***—	—***—	—***—	—***—	—***—
	Girl	Factor					
FSM1	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			-1.5574 (0.6929)		
TruantN	Linear	Linear					
TruantN	No	Factor	—***—	—***—	—***—	—***—	—***—
	Played truant a year ago	Factor					
	1/2	Factor					
	>=3	Factor					

Table B.4.15: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 15)

Variable	Levels	Type	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
AlPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, other relatives	Factor					
	Pros, police	Factor					
	Both	Factor					
AlIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive Media	Factor					
	Interactive Media	Factor					
	Both	Factor					
DgPe1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Parents, other relatives	Factor					
	Pros, police	Factor					
	Both	Factor					
DgIn1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Passive Media	Factor					
	Interactive Media	Factor					
	Both	Factor					
DgEstim	Linear	Linear					
DgEstim	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor			1.052 (0.1319)		
	Half	Factor			2.0303 (0.1758)		
	Most/All	Factor			1.8536 (0.2564)		
Books1	Linear	Linear					-0.8351 (0.2981)
Books1	None	Factor	—***—	—***—	—***—	—***—	—***—
	Few	Factor			0.7698 (0.3398)		
	Lots	Factor			0.874 (0.3209)		
LsDrg	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			-0.3795 (0.1574)		
age	age	Linear			-0.2997 (0.0469)		
gender	Boy	Factor	—***—	—***—	—***—	—***—	—***—
	Girl	Factor			0.2279 (0.0953)		
FSM1	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor				-1.9376 (0.9131)	
TruantN	Linear	Linear				0.3542 (0.1765)	0.4045 (0.188)
TruantN	No	Factor	—***—	—***—	—***—	—***—	—***—
	Played truant a year ago	Factor			0.55 (0.2164)		
	1/2	Factor			0.5799 (0.1676)		
	>=3	Factor			0.757 (0.2148)		

Table B.4.16: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 16)

Variable	Levels	Type	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack
ExclAN1	Linear	Linear					
ExclAN1	No	Factor	—***—	—***—	—***—	—***—	—***—
	Excluded a year ago	Factor					
	1/2	Factor					
	>=3	Factor					
SHA	North East	Factor	—***—	—***—	—***—	—***—	—***—
	North West/Merseyside	Factor					
	Yorkshire	Factor					
	East Midlands	Factor					
	West Midlands	Factor					
	East of England	Factor					
	London	Factor					
	South East Coast	Factor					
	South Central	Factor					
	South West	Factor					

Table B.4.17: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 17)

Variable	Levels	Type	Methadone	Ecstasy	Amphetamines	LSD	Poppers
ExclAN1	Linear	Linear					
ExclAN1	No	Factor	—***—	—***—	—***—	—***—	—***—
	Excluded a year ago	Factor					
	1/2	Factor					
	>=3	Factor					
SHA	North East	Factor	—***—	—***—	—***—	—***—	—***—
	North West/Merseyside	Factor		1.5967 (0.6813)			
	Yorkshire	Factor		0.3855 (0.9338)			
	East Midlands	Factor		1.7678 (0.6855)			
	West Midlands	Factor		0.7876 (0.7891)			
	East of England	Factor		-0.1524 (0.9304)			
	London	Factor		0.5111 (1.1817)			
	South East Coast	Factor		0.8404 (0.6889)			
	South Central	Factor		0.4404 (0.754)			
	South West	Factor		1.8556 (0.6462)			

Table B.4.18: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 18)

Variable	Levels	Type	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
ExclAN1	Linear	Linear				0.4606 (0.2013)	
ExclAN1	No	Factor	—***—	—***—	—***—	—***—	—***—
	Excluded a year ago	Factor					
	1/2	Factor					
	>=3	Factor					
SHA	North East	Factor	—***—	—***—	—***—	—***—	—***—
	North West/Merseyside	Factor			-0.094 (0.2274)		
	Yorkshire	Factor			-0.1999 (0.2485)		
	East Midlands	Factor			-0.005 (0.2148)		
	West Midlands	Factor			-0.3164 (0.2238)		
	East of England	Factor			0.1809 (0.2107)		
	London	Factor			0.5488 (0.2324)		
	South East Coast	Factor			0.3031 (0.2164)		
	South Central	Factor			0.3248 (0.2029)		
	South West	Factor			0.0549 (0.2147)		

Table B.4.19: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 19)

Variable	Levels	Type	Cannabis	Heroin	Cocaine	Magic Mushrooms	Crack
Cannabis	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			1.6668 (0.4986)	0.9395 (0.3176)	1.9017 (0.4439)
Heroin	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			2.551 (0.5783)		2.9283 (0.6214)
Cocaine	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.2865 (0.5155)	2.5125 (0.5984)		1.0955 (0.3521)	2.4939 (0.4957)
Magic Mushrooms	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	0.8624 (0.3430)				1.1872 (0.4791)
Crack	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.2162 (0.5796)	2.7777 (0.6371)	2.2099 (0.5623)		
Methadone	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
Ecstasy	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			2.2053 (0.4229)		
Amphetamines	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			1.5179 (0.4408)	1.3658 (0.3834)	
LSD	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor		1.566 (0.623)		1.5477 (0.4597)	
Poppers	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.6490 (0.2841)		1.5831 (0.3726)		
Ketamine	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.7837 (0.6366)				
Anabolic Steroids	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	2.1927 (0.6126)				
Gas	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	0.4072 (0.1824)	1.8032 (0.4192)		1.1771 (0.2507)	
Other Drugs	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					2.0179 (0.6863)
Tranquillisers	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor				1.6542 (0.5612)	

Table B.4.20: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 20)

Variable	Levels	Type	Methadone	Ecstasy	Amphetamines	LSD	Poppers
Cannabis	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	2.1916 (0.4343)	1.4191 (0.4485)		1.4966 (0.5798)	1.9514 (0.2441)
Heroin	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	2.0821 (0.6208)				
Cocaine	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor		2.2006 (0.4237)	1.6569 (0.4501)		1.6583 (0.3088)
Magic Mushrooms	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor		1.2861 (0.4355)	1.7933 (0.428)	1.6390 (0.4565)	
Crack	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
Methadone	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			1.6259 (0.4546)		
Ecstasy	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.4055 (0.4776)			2.5068 (0.4826)	
Amphetamines	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.5615 (0.4155)	1.6373 (0.4656)			
LSD	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor		2.2788 (0.5645)			
Poppers	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor				0.9975 (0.4465)	
Ketamine	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor		1.5885 (0.5664)	1.6211 (0.615)		
Anabolic Steroids	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.7538 (0.7521)		1.7802 (0.7475)	1.8993 (0.7853)	
Gas	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor		1.1959 (0.3711)			1.1117 (0.2151)
Other Drugs	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
Tranquillisers	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					1.7695 (0.5632)

Table B.4.21: Table of Estimates of Logistic Regression with Smoking, Drinking and Drug-related Socio-demographic Covariates (Table 21)

Variable	Levels	Type	Ketamine	Anabolic Steroids	Gas	Other Drugs	Tranquillisers
Cannabis	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	2.0185 (0.4841)	1.9907 (0.5317)			
Heroin	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	2.2142 (0.6177)		1.1497 (0.4194)	-3.1448 (1.4608)	
Cocaine	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor					
Magic Mushrooms	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			1.0065 (0.2658)	1.2242 (0.5672)	1.5328 (0.6443)
Crack	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor				2.0131 (0.8081)	2.3122 (0.6802)
Methadone	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor		2.1533 (0.6320)			
Ecstasy	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.2595 (0.5049)				2.3826 (0.753)
Amphetamines	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.6013 (0.5032)				
LSD	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	-1.5798 (0.7249)	1.9601 (0.5669)			
Poppers	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			0.9152 (0.2246)		
Ketamine	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor				2.0032 (0.6768)	
Anabolic Steroids	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor			1.7476 (0.4085)	2.6996 (0.8143)	
Gas	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor		1.8619 (0.3803)			
Other Drugs	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	1.4335 (0.6955)				1.8193 (0.8115)
Tranquillisers	No	Factor	—***—	—***—	—***—	—***—	—***—
	Yes	Factor	2.0699 (0.634)				

Appendix C

Results of Log-linear Analysis Models

C.1 Results of Final Log-linear Analysis Model with Backward Elimination

Table C.1.1: Table of Estimates of Log-linear Analysis Final Model (Table 1)

Intercept/Interaction Terms	Complete Case	MICE Scheme 1	MICE Scheme 2
(Intercept)	8.6387 (0.0133)	8.6995 (0.0130)	8.6896 (0.0130)
Cannabis	-2.8730 (0.0559)	-2.9274 (0.0565)	-2.9873 (0.0585)
Heroin	-7.6983 (0.5011)	-6.5619 (0.3532)	-6.2344 (0.3300)
Cocaine	-6.3377 (0.2940)	-6.2065 (0.2779)	-5.8472 (0.2630)
Magic Mushrooms	-5.7263 (0.2159)	-5.5128 (0.1943)	-5.3624 (0.1821)
Crack	-7.2246 (0.4217)	-6.3464 (0.3047)	-6.0955 (0.2916)
Methadone	-6.5206 (0.3340)	-6.3965 (0.3146)	-6.1319 (0.2935)
Ecstasy	-6.8299 (0.3639)	-6.2547 (0.2884)	-6.0214 (0.2746)
Amphetamines	-6.6223 (0.3548)	-6.2799 (0.2995)	-5.9754 (0.2801)
LSD	-7.0481 (0.4485)	-6.8175 (0.4063)	-6.4608 (0.3659)
Poppers	-5.3906 (0.1818)	-5.3589 (0.1770)	-5.2167 (0.1778)
Ketamine	-7.0386 (0.4275)	-6.6562 (0.3529)	-6.4284 (0.3376)
Anabolic Steroids	-7.3812 (0.4461)	-6.8056 (0.3466)	-6.3037 (0.3263)
Gas	-2.6561 (0.0512)	-2.6301 (0.0494)	-2.6102 (0.0496)
Other Drugs	-6.8548 (0.4012)	-6.5778 (0.3490)	-6.1741 (0.3596)
Tranquillisers	-7.7460 (0.5404)	-6.4687 (0.3177)	-6.1323 (0.2933)
Cannabis:Heroin	2.0494 (0.6258)		1.7744 (0.5917)
Cannabis:Cocaine	1.5942 (0.4633)	2.6603 (0.3659)	2.1973 (0.4190)
Cannabis:Magic Mushrooms	2.6415 (0.2988)	2.7483 (0.2446)	2.6743 (0.2549)
Cannabis:Crack	2.4988 (0.5807)	2.5854 (0.4470)	2.5057 (0.4590)
Cannabis:Methadone	2.6032 (0.4658)	3.2091 (0.3957)	3.1357 (0.3972)
Cannabis:Ecstasy	2.8659 (0.4534)	3.1453 (0.3602)	2.5394 (0.4074)
Cannabis:Amphetamines	2.8524 (0.4625)	2.7976 (0.3850)	2.8432 (0.3378)
Cannabis:LSD	2.2947 (0.6867)	3.0956 (0.5194)	2.5151 (0.5652)
Cannabis:Poppers	3.0823 (0.2297)	3.2712 (0.2177)	3.2956 (0.2101)
Cannabis:Ketamine	3.5164 (0.5172)	3.3829 (0.4514)	3.4512 (0.4257)
Cannabis:Anabolic Steroids	1.9593 (0.5649)	2.6927 (0.4275)	2.5370 (0.4847)
Cannabis:Gas	0.8386 (0.1408)	0.8286 (0.1367)	0.9007 (0.1361)
Cannabis:Other Drugs	2.2018 (0.5778)	2.4290 (0.4944)	2.5765 (0.5193)
Cannabis:Tranquillisers	3.9651 (0.8119)	4.5754 (0.5951)	4.2052 (1.1944)
Heroin:Cocaine	5.4590 (0.5874)	4.9954 (0.6540)	3.3535 (0.6928)
Heroin:Magic Mushrooms	10.1242 (0.8481)	8.0758 (1.6748)	4.2922 (1.2889)
Heroin:Crack	-3.2287 (0.8813)		
Heroin:Methadone	9.9682 (1.4740)		
Heroin:Ecstasy			
Heroin:Amphetamines		-5.4199 (1.7716)	-3.7368 (1.4289)

Table C.1.2: Table of Estimates of Log-linear Analysis Final Model (Table 2)

Intercept/Interaction Terms	Complete Case	MICE Scheme 1	MICE Scheme 2
Heroin:LSD	-4.6051 (1.2681)		
Heroin:Poppers			
Heroin:Ketamine	2.7733 (1.3817)		
Heroin:Anabolic Steroids	-16.1859 (1.6611)		
Heroin:Gas	2.9951 (0.5793)	1.9508 (0.5323)	1.3216 (0.5622)
Heroin:Other Drugs	24.8226 (4.0921)		
Heroin:Tranquillisers	8.1571 (1.9761)	-5.3915 (1.7094)	-2.6531 (0.9890)
Cocaine:Magic Mushrooms			
Cocaine:Crack	4.7793 (0.6840)	2.3370 (0.6015)	2.6893 (0.6042)
Cocaine:Methadone	3.0632 (0.7450)		
Cocaine:Ecstasy	2.9912 (0.4206)	2.1424 (0.5031)	2.6359 (0.3950)
Cocaine:Amphetamines	1.3299 (0.5638)	2.8990 (0.4591)	2.1635 (0.5187)
Cocaine:LSD	1.7208 (0.6703)		
Cocaine:Poppers	2.8853 (0.4322)	1.5611 (0.3874)	1.9130 (0.4657)
Cocaine:Ketamine		-7.1597 (2.1158)	
Cocaine:Anabolic Steroids			
Cocaine:Gas			
Cocaine:Other Drugs	3.5488 (0.6827)	2.8327 (0.6787)	
Cocaine:Tranquillisers		3.2159 (1.0143)	
Magic Mushrooms:Crack	2.5952 (0.6377)		
Magic Mushrooms:Methadone			
Magic Mushrooms:Ecstasy	1.2963 (0.5546)		1.5760 (0.5424)
Magic Mushrooms:Amphetamines		3.1545 (0.5008)	2.5612 (0.4685)
Magic Mushrooms:LSD	1.6872 (0.6525)		
Magic Mushrooms:Poppers	1.0592 (0.3552)		
Magic Mushrooms:Ketamine	4.0600 (1.0467)		
Magic Mushrooms:Anabolic Steroids	3.9594 (0.7208)		
Magic Mushrooms:Gas	0.6949 (0.3094)	0.9628 (0.3040)	0.8456 (0.2934)
Magic Mushrooms:Other Drugs	4.3294 (0.6233)	3.3619 (0.6221)	2.6826 (0.6820)
Magic Mushrooms:Tranquillisers	3.5282 (0.7860)		
Crack:Methadone	-7.6057 (1.4261)		
Crack:Ecstasy			
Crack:Amphetamines			
Crack:LSD	-3.6968 (1.1509)		
Crack:Poppers	4.6667 (0.6582)	2.3066 (0.8431)	
Crack:Ketamine	-11.2412 (1.7605)		

Table C.1.3: Table of Estimates of Log-linear Analysis Final Model (Table 3)

Intercept/Interaction Terms	Complete Case	MICE Scheme 1	MICE Scheme 2
Crack:Anabolic Steroids	-36.9802 (3.6192)		
Crack:Gas	1.3836 (0.5917)		
Crack:Other Drugs	5.1968 (1.0235)		3.2040 (1.3345)
Crack:Tranquillisers	2.2393 (1.0872)		
Methadone:Ecstasy	2.8503 (0.6560)		1.6108 (0.5330)
Methadone:Amphetamines	3.0465 (0.7214)	2.9741 (0.5376)	
Methadone:LSD	-11.5027 (1.4053)		
Methadone:Poppers			
Methadone:Ketamine	11.1938 (1.6745)		
Methadone:Anabolic Steroids	5.9007 (1.6878)		2.8328 (1.2345)
Methadone:Gas			
Methadone:Other Drugs			
Methadone:Tranquillisers	-9.4464 (1.5339)		
Ecstasy:Amphetamines			
Ecstasy:LSD	2.9886 (0.7048)	3.7084 (0.5284)	3.3840 (0.5483)
Ecstasy:Poppers			
Ecstasy:Ketamine	-9.4239 (1.7292)	3.3946 (1.0444)	2.5057 (0.5918)
Ecstasy:Anabolic Steroids	3.7481 (0.8926)		
Ecstasy:Gas	1.0204 (0.3739)		
Ecstasy:Other Drugs			
Ecstasy:Tranquillisers			
Amphetamine:LSD	3.2951 (1.0175)		
Amphetamine:Poppers	1.7033 (0.4387)		
Amphetamine:Ketamine			
Amphetamine:Anabolic Steroids	2.9525 (0.8837)		
Amphetamine:Gas			
Amphetamine:Other Drugs			
Amphetamine:Tranquillisers	8.3828 (0.9911)	3.2970 (0.9755)	4.2596 (1.4997)
LSD:Poppers	1.9083 (0.5984)		1.2350 (0.5771)
LSD:Ketamine	-10.2011 (1.6990)		
LSD:Anabolic Steroids	4.8315 (1.0290)		
LSD:Gas			
LSD:Other Drugs			
LSD:Tranquillisers	-6.4586 (1.2021)		
Poppers:Ketamine	2.3247 (0.7824)	3.4330 (1.1524)	
Poppers:Anabolic Steroids	-5.3415 (1.5345)		
Poppers:Gas	0.9826 (0.2338)	1.0958 (0.2230)	0.8350 (0.2262)
Poppers:Other Drugs			
Poppers:Tranquillisers			
Ketamine:Anabolic Steroids			
Ketamine:Gas	1.7688 (0.6846)		
Ketamine:Other Drugs	2.1809 (0.9066)		
Ketamine:Tranquillisers	2.1035 (0.9230)		
Anabolic Steroids:Gas	2.4697 (0.5161)	2.3035 (0.4393)	1.5286 (0.4837)
Anabolic Steroids:Other Drugs	3.6215 (1.0770)		
Anabolic Steroids:Tranquillisers	6.9541 (0.9301)		
Gas:Other Drugs			
Gas:Tranquillisers	2.9826 (0.6418)		
Other Drugs:Tranquillisers	3.0007 (0.9493)		

C.2 Results of Saturated Log-linear Analysis Model

Table C.2.1: Table of Estimates of Log-linear Analysis Saturated Model (Table 1)

Variable	Complete Case	MICE Scheme 1	MICE Scheme 2
(Intercept)	8.6388 (0.0133)	8.7027 (0.0129)	8.6928 (0.0130)
Cannabis	-2.8569 (0.0561)	-2.8387 (0.0558)	-2.8639 (0.0568)
Heroin	-7.6369 (0.5258)	-7.0236 (0.4461)	-6.5012 (0.3936)
Cocaine	-6.3192 (0.2999)	-6.1732 (0.2804)	-5.9011 (0.2791)
Magic Mushrooms	-5.7469 (0.2197)	-5.6533 (0.2105)	-5.4871 (0.2076)
Crack	-7.2169 (0.4350)	-6.6910 (0.3894)	-6.3898 (0.3701)
Methadone	-6.5908 (0.3439)	-6.4875 (0.3253)	-6.2043 (0.2921)
Ecstasy	-6.8840 (0.3751)	-6.4514 (0.3072)	-6.1599 (0.2858)
Amphetamines	-6.6221 (0.3588)	-6.4464 (0.3353)	-6.0856 (0.2850)
LSD	-7.0440 (0.4483)	-6.9156 (0.4228)	-6.5439 (0.3690)
Poppers	-5.3788 (0.1825)	-5.3978 (0.1857)	-5.2889 (0.1770)
Ketamine	-6.9139 (0.4208)	-6.6810 (0.3668)	-6.4742 (0.3491)
Anabolic Steroids	-7.3559 (0.4708)	-6.9773 (0.3980)	-6.4945 (0.3448)
Gas	-2.6587 (0.0514)	-2.6414 (0.0496)	-2.6271 (0.0500)
Other Drugs	-6.8489 (0.4012	-6.9221 (0.4289	-6.4091 (0.3787)
Tranquillisers	-7.8970 (0.6643)	-6.8756 (0.4167)	-6.3464 (0.3352)
Cannabis,Heroin	1.5922 (0.8331)	1.1944 (0.9038)	1.0200 (1.0859)
Cannabis,Cocaine	1.9796 (0.4988)	2.1315 (0.4371)	2.0953 (0.4271)
Cannabis,Magic Mushrooms	2.6575 (0.3080)	2.5363 (0.3189)	2.4602 (0.3204)
Cannabis,Crack	2.0737 (0.6359)	2.3058 (0.6328)	2.3195 (0.5450)
Cannabis,Methadone	2.4135 (0.4928)	2.5520 (0.4732)	2.4646 (0.4781)
Cannabis,Ecstasy	2.8095 (0.4925)	2.4534 (0.4655)	2.3148 (0.4721)
Cannabis,Amphetamines	2.4949 (0.5308)	2.6665 (0.4740)	2.3890 (0.4390)
Cannabis,LSD	2.3539 (0.6931)	2.1373 (0.7802)	2.0860 (0.7669)
Cannabis,Poppers	3.0095 (0.2342)	3.1140 (0.2417)	3.1514 (0.2196)
Cannabis,Ketamine	3.3542 (0.5318)	3.2651 (0.4794)	3.0348 (0.4873)
Cannabis,Anabolic Steroids	1.8757 (0.6584)	1.9102 (0.5764)	1.9359 (0.7501)
Cannabis,Gas	0.8070 (0.1434)	0.7686 (0.1395)	0.7853 (0.1550)
Cannabis,Other Drugs	2.1424 (0.6674)	2.2955 (0.8104)	1.7543 (0.8253)
Cannabis,Tranquillisers	3.6764 (1.2131)	3.4350 (1.7206)	3.1358 (2.6803)
Heroin,Cocaine	5.3682 (0.8476)	4.8479 (0.8198)	3.8530 (1.0391)
Heroin,Magic Mushrooms	8.7213 (1.2899)	7.8793 (2.1346)	4.6862 (3.1461)
Heroin,Crack	-2.6660 (1.1362)	-0.4403 (1.2774)	1.1283 (1.3643)
Heroin,Methadone	10.9827 (1.9091)	5.4948 (2.7236)	2.5665 (2.6411)
Heroin,Ecstasy	0.6014 (1.0584)	-0.0712 (2.0460)	-0.0699 (2.2660)
Heroin,Amphetamines	0.0805 (1.6813)	-3.2297 (2.7543)	-5.3321 (3.6009)

Table C.2.2: Table of Estimates of Log-linear Analysis Saturated Model (Table 2)

Variable	Complete Case	MICE Scheme 1	MICE Scheme 2
Heroin,LSD	-3.6646 (1.9359)	-2.1728 (2.3425)	1.2382 (3.6023)
Heroin,Poppers	0.4906 (1.1034)	-1.1617 (3.9537)	-1.4205 (2.5400)
Heroin,Ketamine	1.7384 (1.8810)	3.5559 (3.9192)	4.4069 (3.5355)
Heroin,Anabolic Steroids	-16.8424 (2.3945)	-7.6430 (5.5743)	-1.4620 (4.7958)
Heroin,Gas	2.9401 (0.6371)	2.2642 (0.7633)	1.9203 (0.6580)
Heroin,Other Drugs	18.7881 (7.0799)	-7.2972 (10.2251)	-9.7594 (9.3113)
Heroin,Tranquillisers	5.9471 (3.2136)	-5.4824 (3.6285)	-3.0579 (5.0752)
Cocaine,Magic Mushrooms	-0.3588 (0.7809)	1.0377 (0.7907)	0.7823 (0.7583)
Cocaine,Crack	4.9842 (0.8129)	3.3908 (0.8925)	2.4443 (1.0605)
Cocaine,Methadone	2.1052 (1.1349)	-0.5190 (2.0081)	-0.3682 (1.8388)
Cocaine,Ecstasy	2.7708 (0.5982)	2.4755 (0.6662)	2.7114 (0.6643)
Cocaine,Amphetamines	1.2360 (0.7633)	1.8509 (0.8073)	2.2651 (1.0601)
Cocaine,LSD	1.0915 (0.8735)	0.6897 (1.1598)	-0.8270 (2.0235)
Cocaine,Poppers	2.7285 (0.5093)	2.0369 (0.4893)	1.8163 (0.4692)
Cocaine,Ketamine	-4.4652 (4.0171)	-12.3839 (4.5758)	-6.6162 (4.9985)
Cocaine,Anabolic Steroids	0.0678 (2.0228)	2.3556 (2.2840)	1.2312 (3.5170)
Cocaine,Gas	-0.6783 (0.5672)	-0.2365 (0.5025)	-0.4986 (0.5834)
Cocaine,Other Drugs	3.4511 (1.3750)	4.4036 (1.6105)	2.2376 (2.1843)
Cocaine,Tranquillisers	3.8069 (1.7850)	3.5487 (1.3525)	1.0234 (2.7542)
Magic Mushrooms,Crack	2.4951 (0.8257)	1.4783 (1.1083)	2.1599 (1.4392)
Magic Mushrooms,Methadone	0.6776 (1.2878)	-0.8019 (1.8617)	-1.0768 (2.0978)
Magic Mushrooms,Ecstasy	0.7556 (0.9061)	1.3482 (0.8390)	1.2912 (0.8671)
Magic Mushrooms,Amphetamines	1.5718 (0.7884)	2.4795 (0.7183)	2.7080 (0.8402)
Magic Mushrooms,LSD	1.3746 (0.7964)	1.2979 (0.9729)	0.8541 (1.0626)
Magic Mushrooms,Poppers	0.9120 (0.4007)	0.4940 (0.4373)	0.3569 (0.4205)
Magic Mushrooms,Ketamine	4.7907 (1.7557)	2.7577 (1.9497)	0.4623 (2.1555)
Magic Mushrooms,Anabolic Steroids	4.1416 (0.8557)	2.2701 (1.6644)	1.7899 (1.6193)
Magic Mushrooms,Gas	0.8403 (0.3299)	0.8970 (0.3563)	0.8899 (0.3633)
Magic Mushrooms,Other Drugs	4.5787 (1.3622)	4.8967 (1.5682)	3.2839 (1.4191)
Magic Mushrooms,Tranquillisers	2.1758 (1.7195)	-0.2789 (2.3706)	1.4175 (3.2168)
Crack,Methadone	-7.3968 (2.0512)	-0.9334 (3.5285)	1.4824 (3.4160)
Crack,Ecstasy	-0.4364 (0.9756)	-0.8509 (2.1362)	-1.8502 (3.0495)
Crack,Amphetamines	0.8858 (1.5161)	-0.9712 (2.4718)	-0.6642 (2.8393)
Crack,LSD	-2.9536 (1.6398)	0.3236 (2.5671)	0.2080 (2.3827)
Crack,Poppers	4.5590 (0.7711)	3.7729 (1.0883)	2.3593 (1.5063)
Crack,Ketamine	-9.0979 (3.8906)	-4.8484 (5.0607)	-2.3669 (3.5224)

Table C.2.3: Table of Estimates of Log-linear Analysis Saturated Model (Table 3)

Variable	Complete Case	MICE Scheme 1	MICE Scheme 2
Crack,Anabolic Steroids	-35.0118 (5.7580)	-9.7267 (8.1413)	-5.9600 (4.5979)
Crack,Gas	1.7930 (0.6273)	0.4910 (0.7488)	0.4594 (0.6705)
Crack,Other Drugs	4.5841 (1.3669)	5.3322 (1.5962)	5.0691 (1.7097)
Crack,Tranquillisers	3.5243 (1.3505)	2.8388 (1.6975)	0.2710 (3.3666)
Methadone,Ecstasy	2.5955 (0.7265)	1.7379 (1.0396)	1.6648 (0.9683)
Methadone,Amphetamines	1.9829 (1.0368)	1.2812 (1.3530)	1.5664 (1.8149)
Methadone,LSD	-11.7509 (2.2931)	-4.3936 (2.9405)	-1.6491 (2.7202)
Methadone,Poppers	1.4693 (0.7081)	1.5173 (0.7199)	0.9016 (0.8934)
Methadone,Ketamine	9.6484 (3.4589)	0.7186 (2.6657)	1.2068 (1.7376)
Methadone,Anabolic Steroids	4.5771 (2.4934)	3.0511 (5.4712)	2.4435 (5.5509)
Methadone,Gas	0.7162 (0.6179)	0.6908 (0.5863)	0.3693 (0.6156)
Methadone,Other Drugs	1.6276 (2.8919)	-1.5581 (3.2643)	0.2218 (2.8647)
Methadone,Tranquillisers	-11.7082 (3.0845)	-2.2900 (3.9049)	-2.1569 (3.9274)
Ecstasy,Amphetamines	0.2941 (0.8462)	1.3720 (0.7215)	0.8630 (0.7289)
Ecstasy,LSD	2.3349 (1.1868)	4.0120 (1.2461)	4.3446 (2.8789)
Ecstasy,Poppers	0.4928 (0.5617)	0.3776 (0.7116)	0.2409 (0.6375)
Ecstasy,Ketamine	-7.2452 (3.4608)	2.4133 (1.9213)	2.1881 (1.6414)
Ecstasy,Anabolic Steroids	4.2572 (1.3517)	0.9766 (2.7171)	-1.2053 (4.3357)
Ecstasy,Gas	1.2271 (0.4284)	0.9976 (0.4440)	0.8602 (0.4587)
Ecstasy,Other Drugs	-0.5450 (1.9387)	-1.1569 (3.2462)	-1.3400 (2.7843)
Ecstasy,Tranquillisers	-0.7987 (1.4463)	0.4081 (1.9182)	-0.1325 (2.8477)
Amphetamine,LSD	3.6428 (1.7912)	-1.7673 (2.3348)	-0.6816 (3.6626)
Amphetamine,Poppers	1.6198 (0.4924)	0.8144 (0.4794)	0.3153 (0.5835)
Amphetamine,Ketamine	-1.3590 (1.4601)	-0.8723 (1.8846)	0.8848 (1.2198)
Amphetamine,Anabolic Steroids	2.9526 (1.0033)	0.9346 (1.7053)	0.9771 (1.8914)
Amphetamine,Gas	0.9857 (0.5120)	0.8473 (0.5507)	0.7539 (0.7232)
Amphetamine,Other Drugs	1.4189 (2.5866)	0.8863 (4.1956)	0.2666 (4.1225)
Amphetamine,Tranquillisers	8.1442 (1.9024)	7.9116 (2.1711)	3.8114 (3.3710)
LSD,Poppers	2.0353 (0.7041)	1.9981 (0.8504)	1.5259 (1.0245)
LSD,Ketamine	-10.8492 (2.9463)	1.7895 (3.5546)	-2.8419 (3.8946)
LSD,Anabolic Steroids	4.8480 (1.0877)	4.0462 (1.5560)	2.9886 (1.8591)
LSD,Gas	-0.2065 (0.9770)	-1.5168 (1.3131)	-1.9522 (1.8986)
LSD,Other Drugs	2.7092 (2.5547)	0.7662 (4.1924)	2.1713 (4.2163)
LSD,Tranquillisers	-3.9620 (2.0194)	-2.3787 (3.1288)	1.2837 (2.3338)
Poppers,Ketamine	3.5444 (1.4155)	2.9193 (2.0111)	1.5425 (1.7286)
Poppers,Anabolic Steroids	-5.3443 (1.9324)	-2.2370 (4.1519)	-0.7084 (5.8453)
Poppers,Gas	1.0187 (0.2539)	1.0335 (0.2491)	0.9728 (0.2465)
Poppers,Other Drugs	-0.3069 (1.1758)	-0.6269 (1.5221)	1.0607 (1.4690)
Poppers,Tranquillisers	-1.0848 (2.0813)	-1.3905 (2.5199)	-0.6313 (2.7722)
Ketamine,Anabolic Steroids	1.8154 (1.4362)	2.2472 (5.2820)	1.9630 (3.1875)
Ketamine,Gas	1.2605 (0.9324)	-1.7388 (2.9289)	-0.6068 (2.3743)
Ketamine,Other Drugs	2.3534 (1.1623)	2.1512 (1.4084)	1.8506 (1.7845)
Ketamine,Tranquillisers	2.3032 (1.2276)	0.7887 (1.7203)	1.0690 (2.3755)
Anabolic Steroids,Gas	2.2228 (0.6005)	2.4333 (0.5696)	1.8237 (0.5285)
Anabolic Steroids,Other Drugs	3.7576 (1.7196)	3.3739 (2.0772)	3.5006 (2.7112)
Anabolic Steroids,Tranquillisers	7.1299 (1.0977)	1.0899 (4.6096)	-0.2395 (4.7769)
Gas,Other Drugs	-0.0335 (1.1287)	-0.0844 (1.2897)	-0.4161 (1.7891)
Gas,Tranquillisers	3.2391 (0.8322)	1.5822 (0.8140)	0.6502 (1.1871)
Other Drugs,Tranquillisers	3.5353 (1.3762)	1.7382 (2.3653)	1.9668 (3.6893)

Appendix D

Item Response Theory Result

D.1 Tables of Estimates of Discrimination and Difficulty Factors in OpenBUGS

Table D.1.1: Table of Estimates of Discrimination Factor with Different Priors (Table 1). For details about priors, please refer to Table 6.3.1.

		α_1		α_2		α_3		α_4		α_5	
		posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank
δ_1	Cannabis	2.869 (0.2068)	4	2.853 (0.1969)	4	2.865 (0.205)	5	2.822 (0.2063)	5	2.875 (0.1931)	4
	Heroin	3.519 (0.4264)	11	3.434 (0.4061)	11	3.342 (0.3951)	11	3.252 (0.3793)	11	3.547 (0.4517)	11
	Cocaine	4.447 (0.4745)	14	4.337 (0.4393)	14	4.198 (0.47)	14	4.067 (0.4241)	14	4.442 (0.5122)	14
	Magic Mushrooms	3.43 (0.3293)	10	3.391 (0.3033)	10	3.31 (0.2882)	10	3.238 (0.2701)	10	3.42 (0.3004)	10
	Crack	3.404 (0.4052)	9	3.323 (0.3714)	9	3.211 (0.3565)	9	3.132 (0.3465)	8	3.384 (0.3518)	9
	Methadone	3.269 (0.3387)	8	3.199 (0.3604)	7	3.098 (0.3355)	7	3.042 (0.2992)	7	3.258 (0.3241)	7
	Ecstasy	4.942 (0.6384)	15	4.836 (0.5954)	15	4.576 (0.5055)	15	4.473 (0.4668)	15	4.94 (0.5786)	15
	Amphetamines	4.024 (0.4495)	13	3.935 (0.4028)	12	3.849 (0.4004)	13	3.751 (0.4168)	13	4.063 (0.4187)	13
	LSD	4.007 (0.4801)	12	3.946 (0.4925)	13	3.742 (0.4282)	12	3.665 (0.4331)	12	4.049 (0.5095)	12
	Poppers	3.257 (0.2351)	7	3.244 (0.2443)	8	3.184 (0.2275)	8	3.147 (0.2242)	9	3.267 (0.263)	8
	Ketamine	2.974 (0.3412)	5	2.921 (0.3153)	5	2.812 (0.284)	4	2.784 (0.3178)	4	2.971 (0.3318)	5
	Anabolic Steroids	2.548 (0.3122)	2	2.531 (0.2709)	2	2.433 (0.283)	2	2.406 (0.2773)	2	2.531 (0.3043)	2
	Gas	1.013 (0.07308)	1	1.014 (0.07113)	1	0.9974 (0.06987)	1	0.9967 (0.08254)	1	1.007 (0.07555)	1
	Other Drugs	2.689 (0.3421)	3	2.617 (0.3226)	3	2.547 (0.2898)	3	2.487 (0.3163)	3	2.626 (0.33)	3
	Tranquillisers	3.244 (0.3914)	6	3.167 (0.4117)	6	3.022 (0.3641)	6	2.964 (0.3713)	6	3.212 (0.3903)	6
δ_2	Cannabis	2.887 (0.1954)	4	2.865 (0.1961)	4	2.831 (0.1956)	4	2.84 (0.1984)	5	2.868 (0.2014)	4
	Heroin	3.594 (0.4611)	11	3.452 (0.4565)	11	3.336 (0.434)	11	3.196 (0.3826)	10	3.566 (0.4685)	11
	Cocaine	4.416 (0.4575)	14	4.299 (0.4656)	14	4.18 (0.4397)	14	4.063 (0.418)	14	4.422 (0.4748)	14
	Magic Mushrooms	3.409 (0.2862)	9	3.39 (0.2893)	10	3.309 (0.3013)	10	3.219 (0.2453)	11	3.431 (0.2953)	10
	Crack	3.419 (0.3849)	10	3.29 (0.3698)	9	3.238 (0.3478)	9	3.121 (0.3339)	8	3.391 (0.3917)	9
	Methadone	3.252 (0.3667)	8	3.208 (0.3467)	8	3.121 (0.3404)	7	3.005 (0.3224)	7	3.267 (0.3771)	7
	Ecstasy	4.87 (0.6028)	15	4.771 (0.5619)	15	4.604 (0.5045)	15	4.456 (0.5035)	15	4.915 (0.582)	15
	Amphetamines	3.989 (0.4381)	12	3.935 (0.4356)	13	3.803 (0.3691)	13	3.7 (0.3995)	13	4.008 (0.3914)	12
	LSD	4.04 (0.4907)	13	3.886 (0.4654)	12	3.782 (0.4423)	12	3.633 (0.3949)	12	4.038 (0.4628)	13
	Poppers	3.25 (0.2493)	7	3.204 (0.2488)	7	3.161 (0.2358)	8	3.134 (0.2421)	9	3.271 (0.262)	8
	Ketamine	2.987 (0.3502)	5	2.918 (0.3066)	5	2.825 (0.3043)	5	2.776 (0.2964)	4	2.934 (0.3393)	5
	Anabolic Steroids	2.565 (0.3201)	2	2.525 (0.2793)	2	2.427 (0.3002)	2	2.359 (0.2921)	2	2.591 (0.3069)	2
	Gas	1.015 (0.07675)	1	1.011 (0.07629)	1	1.004 (0.0752)	1	0.9949 (0.0783)	1	1.013 (0.07438)	1
	Other Drugs	2.619 (0.3076)	3	2.582 (0.3124)	3	2.548 (0.2978)	3	2.451 (0.299)	3	2.66 (0.3029)	3
	Tranquillisers	3.246 (0.4195)	6	3.132 (0.3839)	6	3.059 (0.3528)	6	2.925 (0.3393)	6	3.255 (0.4342)	6

Table D.1.2: Table of Estimates of Discrimination Factor with Different Priors (Table 2). For details about priors, please refer to Table 6.3.1.

		α_6		α_7		α_8		α_9		α_{10}	
		posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank
δ_1	Cannabis	2.894 (0.2019)	4	2.709 (0.1885)	9	2.757 (0.1757)	7	2.857 (0.1957)	4	2.865 (0.1909)	4
	Heroin	3.601 (0.4342)	11	2.697 (0.275)	8	2.888 (0.3198)	9	3.391 (0.3986)	11	3.395 (0.4223)	11
	Cocaine	4.391 (0.4701)	14	3.355 (0.3002)	13	3.67 (0.3367)	14	4.209 (0.435)	14	4.23 (0.4125)	14
	Magic Mushrooms	3.456 (0.3107)	10	2.878 (0.2114)	11	3.004 (0.2617)	11	3.369 (0.2841)	10	3.353 (0.2863)	10
	Crack	3.437 (0.3999)	9	2.682 (0.2593)	7	2.829 (0.3144)	8	3.272 (0.3554)	9	3.269 (0.3704)	9
	Methadone	3.28 (0.3477)	8	2.623 (0.2394)	6	2.735 (0.2741)	6	3.102 (0.3303)	7	3.121 (0.344)	7
	Ecstasy	4.95 (0.5711)	15	3.536 (0.2897)	14	3.957 (0.385)	15	4.643 (0.5089)	15	4.674 (0.5264)	15
	Amphetamines	4.101 (0.4718)	13	3.121 (0.2591)	12	3.344 (0.3319)	13	3.874 (0.3959)	13	3.871 (0.4387)	13
	LSD	4.09 (0.5094)	12	2.994 (0.2862)	15	3.255 (0.3449)	12	3.806 (0.4119)	12	3.816 (0.4103)	12
	Poppers	3.262 (0.2541)	7	2.817 (0.1743)	10	2.943 (0.2093)	10	3.194 (0.2353)	8	3.213 (0.2545)	8
	Ketamine	2.941 (0.3142)	5	2.42 (0.2294)	4	2.524 (0.2486)	4	2.881 (0.3096)	5	2.87 (0.3022)	5
	Anabolic Steroids	2.581 (0.3318)	2	2.122 (0.2142)	2	2.173 (0.2633)	2	2.495 (0.2939)	2	2.456 (0.2794)	2
	Gas	1.008 (0.07558)	1	0.9679 (0.07076)	1	0.9721 (0.0667)	1	1.027 (0.07494)	1	1.011 (0.07581)	1
	Other Drugs	2.694 (0.3254)	3	2.185 (0.2376)	3	2.242 (0.2563)	3	2.613 (0.3138)	3	2.608 (0.3079)	3
	Tranquillisers	3.222 (0.4117)	6	2.528 (0.2859)	5	2.659 (0.3194)	5	3.127 (0.3726)	6	3.109 (0.38)	6
δ_2	Cannabis	2.89 (0.2045)	4	2.739 (0.1834)	9	2.759 (0.1835)	7	2.856 (0.1904)	4	2.843 (0.1963)	4
	Heroin	3.561 (0.4378)	11	2.697 (0.2527)	8	2.882 (0.3284)	9	3.347 (0.4057)	11	3.452 (0.4233)	11
	Cocaine	4.453 (0.4737)	14	3.356 (0.2643)	14	3.667 (0.3185)	14	4.152 (0.4167)	14	4.318 (0.4617)	14
	Magic Mushrooms	3.43 (0.2917)	10	2.878 (0.2208)	11	3.002 (0.2424)	11	3.299 (0.2654)	10	3.359 (0.287)	10
	Crack	3.427 (0.3933)	9	2.654 (0.2397)	6	2.801 (0.2915)	8	3.227 (0.3489)	9	3.267 (0.3542)	9
	Methadone	3.288 (0.34)	8	2.665 (0.2448)	7	2.756 (0.2544)	6	3.134 (0.3094)	7	3.189 (0.3288)	7
	Ecstasy	4.956 (0.5823)	15	3.549 (0.3021)	15	3.951 (0.4101)	15	4.596 (0.4622)	15	4.701 (0.5403)	15
	Amphetamines	4.064 (0.4143)	13	3.101 (0.2672)	13	3.366 (0.3191)	13	3.836 (0.3938)	13	3.909 (0.4293)	13
	LSD	4.024 (0.4894)	12	3.006 (0.2818)	12	3.219 (0.3765)	12	3.779 (0.4647)	12	3.84 (0.4482)	12
	Poppers	3.276 (0.2403)	7	2.832 (0.1871)	10	2.939 (0.2164)	10	3.209 (0.2404)	8	3.19 (0.2347)	8
	Ketamine	3.025 (0.3357)	5	2.427 (0.2139)	4	2.511 (0.2585)	4	2.858 (0.3117)	5	2.884 (0.2991)	5
	Anabolic Steroids	2.583 (0.3166)	2	2.151 (0.2312)	2	2.163 (0.2353)	2	2.489 (0.2972)	2	2.505 (0.2893)	2
	Gas	1.011 (0.07314)	1	0.9614 (0.06845)	1	0.9744 (0.07438)	1	1.021 (0.07756)	1	1.015 (0.07374)	1
	Other Drugs	2.681 (0.3117)	3	2.184 (0.2357)	3	2.255 (0.2726)	3	2.586 (0.2809)	3	2.614 (0.3274)	3
	Tranquillisers	3.235 (0.3689)	6	2.541 (0.268)	5	2.636 (0.304)	5	3.06 (0.3601)	6	3.161 (0.4039)	6

Table D.1.3: Table of Estimates of Discrimination Factor with Different Priors (Table 3). For details about priors, please refer to Table 6.3.1.

		α_{11}		α_{12}		α_{13}		α_{14}	
		posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank
δ_1	Cannabis	2.857 (0.1853)	4	2.884 (0.2022)	5	2.617 (0.1761)	9	2.873 (0.1876)	5
	Heroin	3.612 (0.4715)	11	3.407 (0.4104)	11	2.399 (0.2701)	8	3.405 (0.4206)	11
	Cocaine	4.397 (0.4031)	14	4.315 (0.4325)	14	3.134 (0.2745)	14	4.29 (0.4584)	14
	Magic Mushrooms	3.442 (0.2955)	10	3.353 (0.2735)	10	2.649 (0.1855)	10	3.378 (0.3147)	10
	Crack	3.387 (0.389)	9	3.24 (0.3465)	9	2.381 (0.2503)	7	3.272 (0.3688)	9
	Methadone	3.268 (0.3523)	8	3.164 (0.3308)	7	2.361 (0.2145)	6	3.176 (0.3179)	7
	Ecstasy	4.938 (0.5727)	15	4.743 (0.537)	15	3.283 (0.292)	15	4.752 (0.5496)	15
	Amphetamines	4.057 (0.4656)	13	3.922 (0.4079)	13	2.841 (0.2581)	13	3.896 (0.4118)	13
	LSD	4.036 (0.4871)	12	3.845 (0.4553)	12	2.687 (0.2581)	12	3.859 (0.4542)	12
	Poppers	3.262 (0.2467)	7	3.227 (0.2612)	8	2.67 (0.1827)	11	3.222 (0.2409)	8
	Ketamine	3.011 (0.3436)	5	2.879 (0.3459)	4	2.172 (0.2022)	4	2.853 (0.3206)	4
	Anabolic Steroids	2.593 (0.3082)	2	2.47 (0.2902)	2	1.861 (0.199)	2	2.481 (0.2977)	2
	Gas	1.019 (0.07552)	1	1.001 (0.06644)	1	0.9482 (0.06167)	1	1.007 (0.0776)	1
	Other Drugs	2.682 (0.3131)	3	2.577 (0.3108)	3	1.958 (0.2222)	3	2.58 (0.2961)	3
	Tranquillisers	3.258 (0.4144)	6	3.129 (0.4153)	6	2.236 (0.214)	5	3.138 (0.3894)	6
δ_2	Cannabis	2.871 (0.204)	4	2.841 (0.1935)	4	2.62 (0.1795)	9	2.855 (0.1867)	4
	Heroin	3.55 (0.4473)	11	3.413 (0.383)	11	2.407 (0.272)	8	3.486 (0.4504)	11
	Cocaine	4.451 (0.4563)	14	4.324 (0.4682)	14	3.127 (0.2746)	14	4.325 (0.4332)	14
	Magic Mushrooms	3.451 (0.2834)	10	3.349 (0.2783)	10	2.645 (0.2201)	10	3.35 (0.3137)	10
	Crack	3.401 (0.367)	9	3.259 (0.3913)	9	2.382 (0.2333)	6	3.315 (0.3575)	9
	Methadone	3.252 (0.36)	6	3.168 (0.3084)	7	2.388 (0.2262)	7	3.191 (0.3435)	7
	Ecstasy	4.897 (0.6086)	15	4.76 (0.5577)	15	3.316 (0.3038)	15	4.758 (0.553)	15
	Amphetamines	4.072 (0.4648)	13	3.943 (0.4154)	13	2.846 (0.2572)	13	3.95 (0.4257)	13
	LSD	4.047 (0.4895)	12	3.895 (0.4702)	12	2.683 (0.2778)	12	3.947 (0.5103)	12
	Poppers	3.275 (0.2477)	7	3.211 (0.2549)	8	2.661 (0.184)	11	3.213 (0.2443)	8
	Ketamine	3.011 (0.3245)	5	2.904 (0.3253)	5	2.149 (0.2195)	4	2.889 (0.324)	5
	Anabolic Steroids	2.578 (0.3009)	2	2.528 (0.3208)	2	1.87 (0.2)	2	2.523 (0.3079)	2
	Gas	1.003 (0.0708)	1	1.003 (0.07163)	1	0.9458 (0.06762)	1	0.9984 (0.07315)	1
	Other Drugs	2.664 (0.3195)	3	2.601 (0.3182)	3	1.931 (0.2298)	3	2.61 (0.3106)	3
	Tranquillisers	3.278 (0.4101)	8	3.132 (0.4142)	6	2.202 (0.244)	5	3.178 (0.3931)	6

Table D.1.4: Table of Estimates of Difficulty Factor with Different Priors (Table 1). For details about priors, please refer to Table 6.3.1.

		α_1		α_2		α_3		α_4		α_5	
		posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank
δ_1	Cannabis	1.569 (0.0407)	1	1.569 (0.03857)	1	1.568 (0.04119)	1	1.577 (0.04172)	1	1.566 (0.0372)	1
	Heroin	2.871 (0.1182)	11	2.902 (0.1169)	11	2.936 (0.1178)	11	2.97 (0.1247)	11	2.877 (0.1227)	11
	Cocaine	2.417 (0.06398)	3	2.433 (0.06683)	3	2.454 (0.06884)	3	2.476 (0.06945)	3	2.414 (0.06262)	3
	Magic Mushrooms	2.455 (0.07626)	5	2.466 (0.07425)	5	2.481 (0.07537)	5	2.51 (0.07235)	5	2.453 (0.07159)	5
	Crack	2.822 (0.115)	10	2.845 (0.1115)	10	2.884 (0.1151)	10	2.919 (0.1195)	10	2.817 (0.1024)	10
	Methadone	2.801 (0.1033)	9	2.828 (0.1148)	9	2.86 (0.1095)	9	2.889 (0.1106)	9	2.8 (0.09994)	9
	Ecstasy	2.426 (0.06709)	4	2.436 (0.06398)	4	2.468 (0.06733)	4	2.484 (0.06068)	4	2.423 (0.05827)	4
	Amphetamines	2.568 (0.0761)	6	2.582 (0.0767)	6	2.606 (0.0805)	6	2.631 (0.08498)	6	2.563 (0.07297)	6
	LSD	2.745 (0.09424)	7	2.765 (0.09819)	7	2.806 (0.09764)	7	2.833 (0.103)	8	2.746 (0.09953)	7
	Poppers	2.29 (0.0591)	2	2.294 (0.06131)	2	2.314 (0.05906)	2	2.323 (0.06318)	2	2.286 (0.06234)	2
	Ketamine	2.958 (0.1319)	12	2.98 (0.1392)	12	3.029 (0.1316)	12	3.055 (0.1457)	12	2.958 (0.1255)	12
	Anabolic Steroids	3.226 (0.1889)	15	3.239 (0.1699)	15	3.313 (0.1855)	15	3.333 (0.1947)	15	3.235 (0.1826)	15
	Gas	2.778 (0.1581)	8	2.78 (0.1558)	8	2.812 (0.1527)	8	2.801 (0.1548)	7	2.795 (0.1707)	8
	Other Drugs	3.19 (0.1712)	14	3.227 (0.1679)	14	3.273 (0.1805)	14	3.317 (0.2035)	14	3.217 (0.1804)	14
	Tranquillisers	3.006 (0.1347)	13	3.029 (0.142)	13	3.09 (0.1559)	13	3.113 (0.1557)	13	3.002 (0.1417)	13
δ_2	Cannabis	1.562 (0.03986)	1	1.57 (0.03765)	1	1.573 (0.03934)	1	1.572 (0.03948)	1	1.565 (0.04058)	1
	Heroin	2.859 (0.1174)	11	2.898 (0.1247)	11	2.943 (0.1332)	11	2.985 (0.1227)	11	2.871 (0.1187)	11
	Cocaine	2.415 (0.06381)	3	2.441 (0.06443)	3	2.46 (0.06815)	3	2.479 (0.06433)	3	2.417 (0.06412)	3
	Magic Mushrooms	2.455 (0.0736)	5	2.465 (0.06557)	5	2.489 (0.07116)	5	2.509 (0.06687)	5	2.449 (0.06599)	5
	Crack	2.815 (0.1135)	10	2.853 (0.119)	10	2.877 (0.1085)	10	2.92 (0.112)	10	2.82 (0.1096)	10
	Methadone	2.805 (0.1118)	9	2.824 (0.1035)	9	2.855 (0.1161)	9	2.905 (0.1174)	9	2.805 (0.1132)	9
	Ecstasy	2.426 (0.06447)	4	2.444 (0.06045)	4	2.469 (0.06136)	4	2.491 (0.06344)	4	2.425 (0.06243)	4
	Amphetamines	2.569 (0.07355)	6	2.59 (0.08013)	6	2.615 (0.07741)	6	2.644 (0.07977)	6	2.569 (0.07345)	6
	LSD	2.74 (0.09444)	7	2.774 (0.09618)	7	2.806 (0.1011)	8	2.842 (0.09994)	8	2.741 (0.09145)	7
	Poppers	2.291 (0.06204)	2	2.301 (0.0624)	2	2.316 (0.05947)	2	2.329 (0.06547)	2	2.285 (0.06233)	2
	Ketamine	2.955 (0.1346)	12	2.982 (0.1265)	12	3.03 (0.1404)	12	3.062 (0.1411)	12	2.976 (0.1384)	12
	Anabolic Steroids	3.222 (0.1892)	15	3.238 (0.1747)	15	3.323 (0.2085)	15	3.367 (0.2032)	15	3.202 (0.1751)	15
	Gas	2.775 (0.1671)	8	2.788 (0.1564)	8	2.801 (0.1653)	7	2.825 (0.1795)	7	2.78 (0.1615)	8
	Other Drugs	3.216 (0.1711)	14	3.24 (0.1867)	14	3.269 (0.1767)	14	3.338 (0.1998)	14	3.195 (0.1732)	14
	Tranquillisers	3.0 (0.1413)	13	3.044 (0.1438)	13	3.075 (0.1351)	13	3.128 (0.1424)	13	2.99 (0.1451)	13

Table D.1.5: Table of Estimates of Difficulty Factor with Different Priors (Table 2). For details about priors, please refer to Table 6.3.1.

		α_6		α_7		α_8		α_9		α_{10}	
		posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank
δ_1	Cannabis	1.561 (0.03824)	1	1.6 (0.0413)	1	1.586 (0.03982)	1	1.568 (0.03747)	1	1.566 (0.03775)	1
	Heroin	2.856 (0.1101)	11	3.224 (0.1466)	11	3.13 (0.1323)	11	2.916 (0.1192)	11	2.915 (0.1227)	11
	Cocaine	2.419 (0.06475)	3	2.635 (0.07428)	4	2.56 (0.07157)	3	2.45 (0.06323)	3	2.447 (0.0622)	3
	Magic Mushrooms	2.45 (0.07313)	5	2.63 (0.0752)	3	2.59 (0.08767)	5	2.47 (0.06449)	5	2.475 (0.07205)	5
	Crack	2.811 (0.1046)	10	3.13 (0.1237)	10	3.052 (0.1375)	10	2.866 (0.1131)	10	2.867 (0.1224)	10
	Methadone	2.796 (0.1037)	9	3.09 (0.1191)	9	3.03 (0.1194)	9	2.854 (0.1069)	9	2.849 (0.1138)	9
	Ecstasy	2.421 (0.06339)	4	2.652 (0.06653)	5	2.573 (0.06881)	4	2.458 (0.0623)	4	2.455 (0.05875)	4
	Amphetamines	2.556 (0.07344)	6	2.812 (0.08794)	6	2.739 (0.09066)	6	2.602 (0.07695)	6	2.601 (0.08215)	6
	LSD	2.733 (0.09608)	7	3.065 (0.1112)	8	2.96 (0.1123)	8	2.792 (0.09134)	8	2.788 (0.09529)	8
	Poppers	2.288 (0.0653)	2	2.428 (0.06252)	2	2.385 (0.06593)	2	2.308 (0.06199)	2	2.305 (0.06082)	2
	Ketamine	2.962 (0.1308)	12	3.276 (0.1466)	12	3.195 (0.1438)	12	2.995 (0.1314)	12	3.008 (0.1273)	12
	Anabolic Steroids	3.209 (0.1919)	15	3.594 (0.213)	15	3.528 (0.2264)	15	3.271 (0.1891)	15	3.289 (0.19)	15
	Gas	2.793 (0.1639)	8	2.891 (0.1658)	7	2.87 (0.1553)	7	2.754 (0.1595)	7	2.786 (0.1665)	7
	Other Drugs	3.187 (0.1801)	14	3.559 (0.1978)	14	3.503 (0.2006)	14	3.229 (0.1808)	14	3.243 (0.1775)	14
	Tranquillisers	3.008 (0.1537)	13	3.364 (0.1677)	13	3.281 (0.175)	13	3.044 (0.1363)	13	3.054 (0.1426)	13
δ_2	Cannabis	1.563 (0.03987)	1	1.593 (0.04156)	1	1.588 (0.04139)	1	1.567 (0.03683)	1	1.572 (0.0397)	1
	Heroin	2.864 (0.1198)	11	3.219 (0.1305)	11	3.134 (0.1431)	11	2.93 (0.1159)	11	2.906 (0.1213)	11
	Cocaine	2.413 (0.06406)	3	2.629 (0.06813)	3	2.563 (0.07236)	3	2.454 (0.06694)	3	2.439 (0.06468)	3
	Magic Mushrooms	2.452 (0.06825)	5	2.634 (0.07876)	4	2.587 (0.0794)	5	2.482 (0.06954)	5	2.473 (0.07201)	5
	Crack	2.81 (0.1114)	10	3.138 (0.1205)	10	3.066 (0.1194)	10	2.875 (0.1112)	10	2.863 (0.1116)	10
	Methadone	2.793 (0.09614)	9	3.074 (0.1212)	9	3.021 (0.1106)	9	2.845 (0.1021)	9	2.833 (0.11)	9
	Ecstasy	2.42 (0.05541)	4	2.647 (0.07179)	5	2.58 (0.0711)	4	2.461 (0.05894)	4	2.454 (0.06525)	4
	Amphetamines	2.557 (0.07503)	6	2.813 (0.08996)	6	2.739 (0.08822)	6	2.602 (0.07882)	6	2.594 (0.08029)	6
	LSD	2.739 (0.08825)	7	3.051 (0.1077)	8	2.984 (0.1213)	8	2.801 (0.1085)	8	2.786 (0.09552)	8
	Poppers	2.281 (0.05528)	2	2.42 (0.0659)	2	2.389 (0.06985)	2	2.3 (0.05827)	2	2.306 (0.06421)	2
	Ketamine	2.937 (0.1274)	12	3.262 (0.1382)	12	3.203 (0.1455)	12	3.017 (0.1393)	12	3.001 (0.1237)	12
	Anabolic Steroids	3.209 (0.1809)	15	3.562 (0.2107)	14	3.537 (0.2021)	15	3.27 (0.1882)	15	3.258 (0.173)	15
	Gas	2.789 (0.1547)	8	2.903 (0.1592)	7	2.875 (0.1706)	7	2.769 (0.163)	7	2.78 (0.1501)	7
	Other Drugs	3.178 (0.1616)	14	3.567 (0.2137)	15	3.493 (0.2081)	14	3.244 (0.1691)	14	3.239 (0.1886)	14
	Tranquillisers	3.001 (0.1328)	13	3.36 (0.1574)	13	3.292 (0.166)	13	3.067 (0.1503)	13	3.039 (0.1394)	13

Table D.1.6: Table of Estimates of Difficulty Factor with Different Priors (Table 3). For details about priors, please refer to Table 6.3.1.

		α_{11}		α_{12}		α_{13}		α_{14}	
		posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank
δ_1	Cannabis	1.567 (0.03838)	1	1.566 (0.03848)	1	1.617 (0.04237)	1	1.566 (0.03887)	1
	Heroin	2.859 (0.1187)	11	2.911 (0.1213)	11	3.414 (0.1673)	11	2.916 (0.1261)	11
	Cocaine	2.416 (0.05752)	3	2.439 (0.06179)	3	2.71 (0.08446)	3	2.438 (0.0643)	3
	Magic Mushrooms	2.449 (0.07062)	5	2.473 (0.06928)	5	2.728 (0.0835)	4	2.47 (0.07419)	5
	Crack	2.825 (0.1093)	10	2.873 (0.1144)	10	3.311 (0.1588)	10	2.835 (0.1169)	9
	Methadone	2.802 (0.1061)	9	2.837 (0.1023)	9	3.249 (0.1422)	9	2.86 (0.1169)	10
	Ecstasy	2.421 (0.06027)	4	2.445 (0.059)	4	2.732 (0.07888)	5	2.442 (0.06079)	4
	Amphetamines	2.563 (0.08025)	6	2.587 (0.07456)	6	2.925 (0.1021)	6	2.593 (0.08159)	6
	LSD	2.738 (0.09265)	7	2.78 (0.09983)	7	3.207 (0.1232)	8	2.779 (0.1027)	7
	Poppers	2.291 (0.06087)	2	2.303 (0.06195)	2	2.488 (0.0704)	2	2.298 (0.06335)	2
	Ketamine	2.944 (0.1266)	12	3.004 (0.1456)	12	3.463 (0.1567)	12	3.005 (0.14)	12
	Anabolic Steroids	3.208 (0.1752)	15	3.275 (0.1955)	15	3.868 (0.2549)	15	3.268 (0.1917)	15
	Gas	2.772 (0.1629)	8	2.805 (0.1508)	8	2.93 (0.1571)	7	2.798 (0.1675)	8
	Other Drugs	3.192 (0.1688)	14	3.251 (0.1786)	14	3.795 (0.2376)	14	3.239 (0.1803)	14
	Tranquillisers	2.99 (0.1346)	13	3.046 (0.1455)	13	3.578 (0.1718)	13	3.038 (0.1375)	13
δ_2	Cannabis	1.567 (0.03823)	1	1.572 (0.04011)	1	1.617 (0.04416)	1	1.571 (0.03857)	1
	Heroin	2.874 (0.1222)	11	2.909 (0.1135)	11	3.408 (0.1828)	11	2.891 (0.1222)	11
	Cocaine	2.418 (0.0643)	3	2.437 (0.06507)	3	2.712 (0.0786)	3	2.436 (0.06426)	3
	Magic Mushrooms	2.447 (0.06928)	5	2.475 (0.06995)	5	2.73 (0.08575)	4	2.476 (0.0759)	5
	Crack	2.817 (0.1062)	10	2.868 (0.1236)	10	3.31 (0.1556)	10	2.845 (0.1098)	10
	Methadone	2.81 (0.1115)	9	2.834 (0.0986)	9	3.239 (0.1335)	9	2.828 (0.1053)	9
	Ecstasy	2.426 (0.06529)	4	2.447 (0.06131)	4	2.731 (0.08162)	5	2.443 (0.06421)	4
	Amphetamines	2.562 (0.08129)	6	2.592 (0.07941)	6	2.918 (0.1013)	6	2.591 (0.07565)	6
	LSD	2.745 (0.09568)	7	2.779 (0.09834)	7	3.207 (0.1326)	8	2.766 (0.1045)	7
	Poppers	2.289 (0.05936)	2	2.303 (0.06107)	2	2.489 (0.07015)	2	2.302 (0.05845)	2
	Ketamine	2.949 (0.1253)	12	2.997 (0.1307)	12	3.482 (0.1839)	12	2.995 (0.1308)	12
	Anabolic Steroids	3.214 (0.1711)	15	3.251 (0.1818)	15	3.868 (0.238)	15	3.249 (0.1855)	15
	Gas	2.801 (0.1546)	8	2.802 (0.1594)	8	2.945 (0.1671)	7	2.815 (0.1618)	8
	Other Drugs	3.193 (0.1715)	14	3.239 (0.1832)	14	3.839 (0.2635)	14	3.236 (0.1783)	14
	Tranquillisers	2.994 (0.1309)	13	3.054 (0.1479)	13	3.61 (0.1942)	13	3.032 (0.1334)	13

D.2 Result Table for Two-parameter IRT Models under `ltm` and `OpenBUGS`

Table D.2.1: Table of Estimates for Discrimination Factors and Difficulty Factors for Two-parameter IRT Model under `ltm` and `OpenBUGS`

	Discrimination Factor				Difficulty Factor			
	ltm		OpenBUGS		ltm		OpenBUGS	
	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank	posterior estimate(sd)	Rank
Cannabis	2.8699 (0.1943)	4	2.881 (0.1985)	4	1.5493 (0.0387)	1	1.565 (0.0388)	1
Heroin	3.4178 (0.3966)	10	3.47 (0.4098)	11	2.8364 (0.1160)	11	2.893 (0.1143)	11
Cocaine	4.2266 (0.4675)	14	4.316 (0.4427)	14	2.3962 (0.0721)	3	2.433 (0.06171)	3
Magic Mushrooms	3.5408 (0.3390)	11	3.387 (0.2913)	10	2.4013 (0.0736)	4	2.464 (0.06979)	5
Crack	3.3103 (0.3587)	9	3.327 (0.3763)	9	2.7890 (0.1047)	10	2.844 (0.1129)	10
Methadone	3.2548 (0.3429)	6	3.2 (0.3372)	7	2.7663 (0.1055)	9	2.825 (0.1059)	9
Ecstasy	4.3968 (0.4819)	15	4.773 (0.5511)	15	2.4242 (0.0703)	5	2.441 (0.06054)	4
Amphetamines	3.9083 (0.4465)	13	3.953 (0.4234)	13	2.5455 (0.0879)	6	2.582 (0.07853)	6
LSD	3.7161 (0.4177)	12	3.918 (0.4698)	12	2.7584 (0.1000)	8	2.767 (0.09496)	7
Poppers	3.2619 (0.2507)	7	3.228 (0.2466)	8	2.2605 (0.0597)	2	2.297 (0.05977)	2
Ketamine	2.9295 (0.3343)	5	2.925 (0.3268)	5	2.9291 (0.1396)	12	2.981 (0.1357)	12
Anabolic Steroids	2.6123 (0.3012)	3	2.504 (0.3005)	2	3.1431 (0.1636)	14	3.26 (0.1893)	15
Gas	1.0426 (0.0761)	1	1.009 (0.06806)	1	2.6984 (0.1522)	7	2.789 (0.1543)	8
Other Drugs	2.6027 (0.3056)	2	2.62 (0.3139)	3	3.1853 (0.1709)	15	3.223 (0.1784)	14
Tranquillisers	3.2847 (0.4049)	8	3.163 (0.3922)	6	2.9463 (0.1312)	13	3.032 (0.1426)	13

Appendix E

Table of Latent Class Analysis

E.1 Frequency Table of Drug-trying Response Variables in Each Imputed Data Set

Table E.1.1: Frequency Table of Drug-trying Response Variables in Each Imputed Data Set for the `R` and `Latent Gold` programs (Latent Class Analysis model). `R`: `R` program; `G`: `Latent Gold` program

Response	Data Set	1	2	3	4	5	6	7	8	9	10
Cannabis	R	676	669	676	678	678	681	677	680	678	676
	LG	676	681	683	681	675	678	678	676	675	677
Heroin	R	46	44	43	44	44	44	41	43	46	41
	LG	46	44	43	44	44	44	41	43	46	41
Cocaine	R	96	92	94	94	95	96	92	96	97	94
	LG	91	91	92	91	92	92	91	89	94	90
Magic Mushrooms	R	114	114	115	113	116	115	112	113	115	114
	LG	110	113	112	111	112	116	112	111	111	111
Crack	R	52	48	52	50	52	54	49	53	55	50
	LG	47	46	49	45	48	48	47	46	50	48
Methadone	R	56	54	56	55	59	57	53	56	57	56
	LG	53	53	54	54	55	58	54	52	58	54
Ecstasy	R	87	82	84	84	86	87	81	85	87	85
	LG	83	82	83	82	82	82	83	83	83	80
Amphetamines	R	71	70	73	71	75	73	70	73	74	71
	LG	70	71	69	69	68	69	69	70	68	68
LSD	R	46	43	47	45	48	46	45	45	48	46
	LG	44	44	43	43	45	45	44	43	44	47
Poppers	R	170	167	172	169	174	169	171	172	170	172
	LG	167	168	169	170	168	169	169	167	170	169
Ketamine	R	50	45	50	47	49	49	44	46	47	45
	LG	43	45	44	44	46	46	44	44	47	44
Anabolic Steroids	R	38	37	37	38	39	38	38	38	41	39
	LG	38	35	36	35	36	37	35	35	37	36
Gas	R	601	604	609	602	606	606	599	605	601	609
	LG	607	606	604	603	595	610	596	607	603	600
Other Drugs	R	35	34	37	34	33	36	36	35	36	36
	LG	36	35	35	35	36	36	34	33	34	34
Tranquillisers	R	37	33	39	35	34	35	37	35	36	37
	LG	33	34	34	33	34	35	34	36	37	34

Appendix F

Tables of Weighted Results

F.1 Design Factor Table on Five Perspectives of the Year 2010 Study

Table F.1.1: True Standard Error and Design Factor Table on Five Perspectives of the Year 2010 Study.

Key Variables	Gender	Sample Size	Weighted Sample Size	True Standard Errors	Design Factors
Prevalence of regular smoking	Male	3663	3676	0.378	1.166
	Female	3591	3575	0.445	1.083
Proportion who drank alcohol in the last week	Male	3531	3541	0.73	1.292
	Female	3486	3468	0.652	1.083
Mean alcohol consumption in the last week	Male	389	377	0.836	1.169
	Female	401	394	0.986	1.161
Proportion who have taken drugs in the last month	Male	3383	3395	0.556	1.23
	Female	3410	3388	0.471	1.168
Proportion who have taken drugs in the last year	Male	3401	3416	0.67	1.163
	Female	3424	3404	0.657	1.083

F.2 Estimate Tables For Weighted Results

Table F.2.1: Estimate Table of Logistic Regrssion among 15 Drug-trying Response Variables Only (Unweighted model vs weighted model) (Table 1)

Cannabis				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-2.7937	0.0544	-2.7311	0.0707
Cannabis				
Heroin				
Cocaine	1.9203	0.4247	1.4227	0.5797
Magic Mushrooms	1.7261	0.3154	1.9892	0.4274
Crack	1.3056	0.5562	1.6481	0.6776
Methadone	2.2028	0.5141	2.3349	0.6803
Ecstasy	2.1544	0.5094	2.2548	0.6767
Amphetamine	1.1700	0.5102	1.2332	0.7146
LSD	1.9235	0.6816	1.4654	0.9238
Poppers	2.9510	0.2274	3.0144	0.2877
Ketamine	3.0345	0.4917	2.8347	0.5912
Anabolic Steroids	1.6713	0.5272	1.9280	0.6719
Gas	0.6926	0.1439	0.5502	0.1942
Other Drugs	1.5111	0.6388	1.7356	0.8111
Tranquillisers				
Heroin				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.9025	0.3635	-7.0504	0.3313
Cannabis				
Heroin				
Cocaine	3.1359	0.5140	3.8005	0.4288
Magic Mushrooms	1.1325	0.5499		
Crack	2.6261	0.5796	2.5595	0.5080
Methadone				
Ecstasy				
Amphetamine				
LSD				
Poppers				
Ketamine				
Anabolic Steroids	2.0174	0.7362	2.6668	0.6004
Gas	2.0131	0.4535	1.9512	0.3933
Other Drugs				
Tranquillisers				

Table F.2.2: Estimate Table of Logistic Regrssion among 15 Drug-trying Response Variables Only (Unweighted model vs weighted model) (Table 2)

Cocaine				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.2757	0.2793	-6.2124	0.2232
Cannabis	2.4658	0.3672	2.2158	0.3043
Heroin	2.8845	0.5881	3.0657	0.4898
Cocaine				
Magic Mushrooms				
Crack	2.2258	0.5846	2.3501	0.4808
Methadone				
Ecstasy	2.1457	0.3905	2.2510	0.3269
Amphetamine	1.3670	0.4369	1.2733	0.3720
LSD				
Poppers	1.5163	0.3495	1.6392	0.2902
Ketamine				
Anabolic Steroids				
Gas				
Other Drugs	1.5313	0.5970	1.4439	0.5105
Tranquillisers				
Magic Mushrooms				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-5.5673	0.1961	-5.6949	0.1737
Cannabis	2.4120	0.2594	2.4757	0.2330
Heroin	1.3508	0.5966	1.4980	0.4846
Cocaine				
Magic Mushrooms				
Crack	1.2651	0.5339		
Methadone				
Ecstasy			0.8523	0.3456
Amphetamine	1.8743	0.3693	1.2443	0.3435
LSD	1.9126	0.4486	1.3565	0.4390
Poppers			0.7052	0.2653
Ketamine				
Anabolic Steroids				
Gas	1.1088	0.2611	1.0012	0.2243
Other Drugs	1.6800	0.5097		
Tranquillisers			1.4339	0.5487

Table F.2.3: Estimate Table of Logistic Regrssion among 15 Drug-trying Response Variables Only (Unweighted model vs weighted model) (Table 3)

Crack				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.6165	0.3327	-6.4595	0.2830
Cannabis	2.1625	0.4487	2.1587	0.3906
Heroin	2.7471	0.5822	1.9960	0.4927
Cocaine	2.0928	0.4817	2.3254	0.4183
Magic Mushrooms	1.3383	0.4843	1.3361	0.4097
Crack				
Methadone				
Ecstasy				
Amphetamine				
LSD				
Poppers				
Ketamine				
Anabolic Steroids	-1.9414	1.0203		
Gas				
Other Drugs				
Tranquillisers	1.5110	0.7260		
Methadone				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.5812	0.3353	-6.7766	0.3235
Cannabis	2.7722	0.4268	2.7941	0.3940
Heroin	1.6089	0.6077	1.7721	0.5276
Cocaine				
Magic Mushrooms				
Crack				
Methadone				
Ecstasy	1.6518	0.4626	1.4421	0.4173
Amphetamine	1.9609	0.4595	2.0256	0.4101
LSD				
Poppers				
Ketamine				
Anabolic Steroids				
Gas			0.7901	0.3456
Other Drugs				
Tranquillisers				

Table F.2.4: Estimate Table of Logistic Regrssion among 15 Drug-trying Response Variables Only (Unweighted model vs weighted model) (Table 4)

Ecstasy				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-7.0491	0.3895	-7.0659	0.3527
Cannabis	2.9169	0.4501	2.9460	0.4071
Heroin				
Cocaine	2.5103	0.3882	2.5027	0.3482
Magic Mushrooms	0.9971	0.4252	1.0908	0.3843
Crack				
Methadone	1.4226	0.5555	1.3940	0.4727
Ecstasy				
Amphetamine	1.2128	0.4688	1.4593	0.4186
LSD	2.2862	0.5259	2.3957	0.4948
Poppers				
Ketamine	1.7167	0.6530	1.7392	0.5690
Anabolic Steroids				
Gas	1.2246	0.3544	1.0519	0.3256
Other Drugs				
Tranquillisers				
Amphetamine				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.4145	0.3043	-6.5207	0.2362
Cannabis	2.3824	0.4165	2.4432	0.3157
Heroin				
Cocaine	1.1801	0.4338	1.1304	0.3204
Magic Mushrooms	1.6172	0.3912	1.3087	0.3015
Crack				
Methadone	1.7816	0.4862	2.0364	0.3351
Ecstasy	0.9449	0.4486	1.2445	0.3265
Amphetamine				
LSD				
Poppers	1.0854	0.3828	0.9562	0.2854
Ketamine				
Anabolic Steroids				
Gas				
Other Drugs				
Tranquillisers				

Table F.2.5: Estimate Table of Logistic Regrssion among 15 Drug-trying Response Variables Only (Unweighted model vs weighted model) (Table 5)

LSD				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-7.2787	0.4565	-7.0961	0.4137
Cannabis	2.5932	0.5785	2.1728	0.5587
Heroin	1.8542	0.6248	1.9715	0.6015
Cocaine				
Magic Mushrooms	1.9545	0.4442	1.8976	0.4605
Crack				
Methadone				
Ecstasy	2.3914	0.4538	2.5057	0.4708
Amphetamine				
LSD				
Poppers	1.3871	0.4433	1.3561	0.4586
Ketamine	-2.1615	0.8623	-1.8234	0.8098
Anabolic Steroids				
Gas				
Other Drugs				
Tranquillisers				
Poppers				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-5.3490	0.1772	-5.2558	0.1580
Cannabis	3.1558	0.2163	3.1298	0.1927
Heroin				
Cocaine	1.5417	0.3074	1.5223	0.2894
Magic Mushrooms	0.6800	0.3155	0.7702	0.2897
Crack				
Methadone				
Ecstasy				
Amphetamine	1.0371	0.3578	0.8789	0.3375
LSD	0.9255	0.4375	0.9683	0.4269
Poppers				
Ketamine				
Anabolic Steroids				
Gas	0.9871	0.2175	0.9732	0.1981
Other Drugs				
Tranquillisers				

Table F.2.6: Estimate Table of Logistic Regrssion among 15 Drug-trying Response Variables Only (Unweighted model vs weighted model) (Table 6)

Ketamine				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.8171	0.3786	-6.5339	0.3118
Cannabis	3.0986	0.4600	2.8089	0.3937
Heroin				
Cocaine				
Magic Mushrooms				
Crack				
Methadone				
Ecstasy				
Amphetamine	1.8140	0.4737	2.1551	0.4192
LSD				
Poppers				
Ketamine				
Anabolic Steroids				
Gas				
Other Drugs	1.6370	0.6211	1.5236	0.5902
Tranquillisers	2.1775	0.6089	2.3566	0.5758
Anabolic Steroids				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.8057	0.3482	-6.9418	0.3224
Cannabis	2.0530	0.4476	2.1651	0.3850
Heroin	1.8887	0.5867	2.1805	0.4945
Cocaine				
Magic Mushrooms				
Crack				
Methadone				
Ecstasy				
Amphetamine	1.2496	0.5370		
LSD				
Poppers				
Ketamine				
Anabolic Steroids				
Gas	1.8589	0.4106	2.2036	0.3572
Other Drugs	1.6970	0.6703	2.1094	0.5356
Tranquillisers				

Table F.2.7: Estimate Table of Logistic Regrssion among 15 Drug-trying Response Variables Only (Unweighted model vs weighted model) (Table 7)

Gas				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-2.6340	0.0507	-2.6099	0.0502
Cannabis	0.8094	0.1360	0.6881	0.1362
Heroin	1.7214	0.4693	1.1576	0.4695
Cocaine	-0.7606	0.3747		
Magic Mushrooms	0.8989	0.2616	0.8876	0.2678
Crack				
Methadone				
Ecstasy	0.9273	0.3357	0.6596	0.3108
Amphetamine				
LSD				
Poppers	0.9337	0.2224	0.8763	0.2144
Ketamine				
Anabolic Steroids	1.7056	0.4293	2.1136	0.4166
Gas				
Other Drugs				
Tranquillisers				
Other Drugs				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.6843	0.3541	-6.7536	0.3322
Cannabis	2.3636	0.4935	2.4396	0.4476
Heroin	-2.0610	1.1018	-2.2928	0.9541
Cocaine	1.6333	0.5405	1.5320	0.5050
Magic Mushrooms	1.3862	0.5324	1.2402	0.4993
Crack				
Methadone				
Ecstasy				
Amphetamine				
LSD				
Poppers				
Ketamine	1.9220	0.5963	1.9641	0.5286
Anabolic Steroids	1.6530	0.7591	2.5319	0.5942
Gas				
Other Drugs				
Tranquillisers				

Table F.2.8: Estimate Table of Logistic Regrssion among 15 Drug-trying Response Variables Only (Unweighted model vs weighted model) (Table 8)

Tranquillisrs				
	Unweighted		Weighted	
	Estimate	SE	Estimate	SE
(Intercept)	-6.6144	0.3370	-6.8317	0.3679
Cannabis	1.7768	0.5199	1.7215	0.5850
Heroin				
Cocaine				
Magic Mushrooms	1.9219	0.5566	1.5473	0.5993
Crack	1.6722	0.6505		
Methadone				
Ecstasy			2.2648	0.6376
Amphetamine				
LSD				
Poppers				
Ketamine	1.8592	0.6546	1.6527	0.6159
Anabolic Steroids				
Gas				
Other Drugs	1.6866	0.6835		
Tranquillisers				