The Application of Causal Models in the Analysis of Grade 12 Results in Gauteng and Western Cape Provinces

Maupi Eric Letsoalo

A thesis submitted in fulfilment of the requirements for the degree of
DOCTOR OF EDUCATION
in MATHEMATICS EDUCATION

in the

Department of Mathematics, Science and Technology Education

in the
FACULTY OF HUMANITIES
(School of Education)

at the

UNIVERSITY OF LIMPOPO

Supervisor	: **Dr RS Maoto**
Co-supervisor	: Dr JK Masha
Co-supervisor	: Prof M Lesaoana

2016

Declaration

*****

I declare that the thesis hereby submitted to the University of Limpopo, for the degree of Doctor of Education in Mathematics Education has not previously been submitted by me for a degree at this or any other university; that it is is my work in design and in execution, and that all material contained herein has been duly acknowledged.

Letsoalo

16th August 2016

Letsoalo, ME (Mr.)

Date

Dedication

*****

My daughter Mampe, my son Maupe, my brothers Maboke and Pompi, my sister Jelane and my parents Mmambi and Alfred

Ngoakoana Ester Khomo

The late Johannes Ramabele Maibelo. Ngwato!

Acknowledgements

*****

...somewhere in the middle of this project - I had realised that I was in the process of learning...I was beginning to learn. When I got to the end of the project - the realisation that I (still) have a lot to learn had cristallised in me.

Everyone has a journey uniquely their own to find personal truths, and this thesis reflects part of my personal academic journey. The academic journey is not easy; as such one needs the support of other willing-individuals to complete it. This project is no different as I would love to thank some individuals whom without their support and willingness to contribute I would not have completed it.

First, I would like to express my thanks to my Creator, Almighty God, for His mercy and for His blessings, and for making everything possible.

My supervisors, Dr. RS Maoto, Dr. JK Masha and Prof. M Lesaoana, for their critical input, and without whose exceptional wisdom, promptness, patience, and encouragement (intellectual guidance) this study would certainly not have been possible;

Special thanks to my wife Mamosana

“Tlogolo sa Tau tša ga-Tshebetshebe;
Ngwana’ Tlou tša ga-Makgoba”,

for her unwavering support and encouragement, and for her famous expression:

“. . . but you will finish . . . It is just another journey”,

My brother, Matome David Letsoalo,

”La ’Matau’a Magaša; Matome tha’ta khê-epe, a rema ga tee a lesa;
Bare Khê-Jeêjeê . . . khomu-monarene!
Nkadime leselo ke fefere, bana baka ba letše ka tlala maleng”,

for proper guidance, motivation, encouragement and inspiration;

Tshwane University of Technology’s Information Librarian, Ms Eunice Mtshali, for some technical assistance, and

Umalusi for allowing me to use their privileged data; especially Dr MS Rakometsi (for making it possible for me to access Grade 12 dataset), Mr E Sibanda (for facilitation of the process of getting Grade 12 dataset) and Ms LP Masemola (for data management at SIR: Umalusi).

Abstract

*****

The focus in this thesis was on the approaches that seek to compare two study arms in the absence of randomisation when the interclass correlation coefficient is greater than zero. Many reports on performance of learners in Grade 12 have used ordinary regression models (such as logistic regression model and linear regression models) which ignore clustering effect, and descriptive statistics (e.g., averages and standard deviations for continuous variables, and proportions expressed as percentages and frequencies). These models do not only bias point estimates but also give falsely narrow confidence intervals. The study was applied to two of the nine provinces of South Africa: Gauteng Province and Western Cape Province in 2008, 2009 and 2010 academic years.

Causal models, and in particular, hierarchical models (or disaggregated approach), unlike descriptive analyses, are more powerful as they are able to adjust for individual covariates. For the analysis of continuous variables; Western Cape Province was expected to significantly score higher marks than Gauteng Province in 2008 (Crude estimate: 0.782) and 2009 (Crude estimate: 0.957) while Gauteng Province was expected to score higher marks than Western Cape Province in 2010 (Crude estimate: -0.302). Adjusted models indicate that Western Cape Province performed better than Gauteng Province in 2008 and 2009 but not in 2010 where Gauteng Province performed better than Western Cape Province after adjusting for gender. In case of binary outcome; the crude estimates favoured Western Cape Province than Gauteng Province in 2008 (Odds ratio = 1.16) and 2009 (Odds ratio = 1.19). Otherwise, the crude estimates favoured Gauteng Province in 2010 (Odds ratio = 0.11).

The proportion of female learners in Gauteng Province ranged between 54.48% and 54.99%, while in Western Cape Province it ranged between 56.78% and 57.16%, in 2008 through 2010 academic years. Proportion of female learners in Western Cape Province were found to be higher than those in Gauteng during this period. At least 70.42% of learners in Gauteng and at least 73.96% of learners in Western Cape Province passed Grade 12 during the years 2008 to 2010.

Through the application of causal model we have learned that although gender is not a significant predictor of the overall learner performance in Grade 12, the effect of gender gave the mixed findings depending on the nature of the outcome. The effect of gender on continuous endpoint (marks) suggests that a model of single-sex classrooms or single-sex schools may be adopted so as to mitigate the inherent perceptions and stereotype regarding learner-gender. However, the results based on binary endpoint (pass/not pass) suggest that coeducation system is the best bet.

A school quintile is a significant predictor of the overall learner performance in the two provinces. The resourceful schools are more likely to produce learners with higher marks. Also, the resourced schools than the less or under resourced schools are more likely to produce the favourable results (higher marks (%) or/and pass) in the two provinces.

Key words: Hierarchical models, Learner performance, Nonrandomisation, Intra-class correlation coefficient, School quintile, Single-sex education

List of Acronyms

*****

ACE	Adverse childhood experience
AHRB	Adoption of health risk behaviour
ANCOVA	Analysis of Covariates
ATE	Average Treatment Effect
CART	Classification and regression tree analysis
CRD	Completely randomised experiment
DBE	Department of Basic Education
DDSP	Disease, disability and social problems
DETE	Department of Education and Employment
DHET	Department of Higher Education and Training
EPBR	Equal-percent bias reducing
FET	Further Education and Training
FM	Full Matching
GEE	Generalised estimating equation
GFET	General and Further Education and Training
GLL	Generalised log-likelihood
GP	Gauteng Province
HC	Higher Certificate
ICC	Intracluster Correlation Coefficient
LD	Learning disability
LME	Linear mixed-effects model

MECs	Members of executive council
MES	Ministry of Education Singapore
NDoE	National Department of Education
NID	Normally and Independently Distributed
NNM	Nearest Neighbour Matching
NLME	Nonlinear mixed-effects model
NPME	Nonparametric mixed-effects
NQF	National Qualifications Framework
NSC	National Senior Certificate
OTL	Opportunity to Learn
OR	Odds Ratio
PSM	Propensity Score Matching
QA	Quality Assurance
RCT	Randomised Controlled Trial
RIE	Randomised Impact Evaluations
RSA	Republic of South Africa
SC	Senior Certificate
SIP	School Improvement Plan
SECI	Social, emotional, and cognitive impairment
SES	Socio-Economic Status
SUTVA	Stable unit treatment value assumption
TUT	Tshwane University of Technology
WCP	Western Cape Province

Contents

*****

List of Figures

*****

List of Tables

*****

Chapter 1

Orientation to the Study

*****

Absenteeism is a significant problem at many institutions of (higher) learning.

Romer (1993)

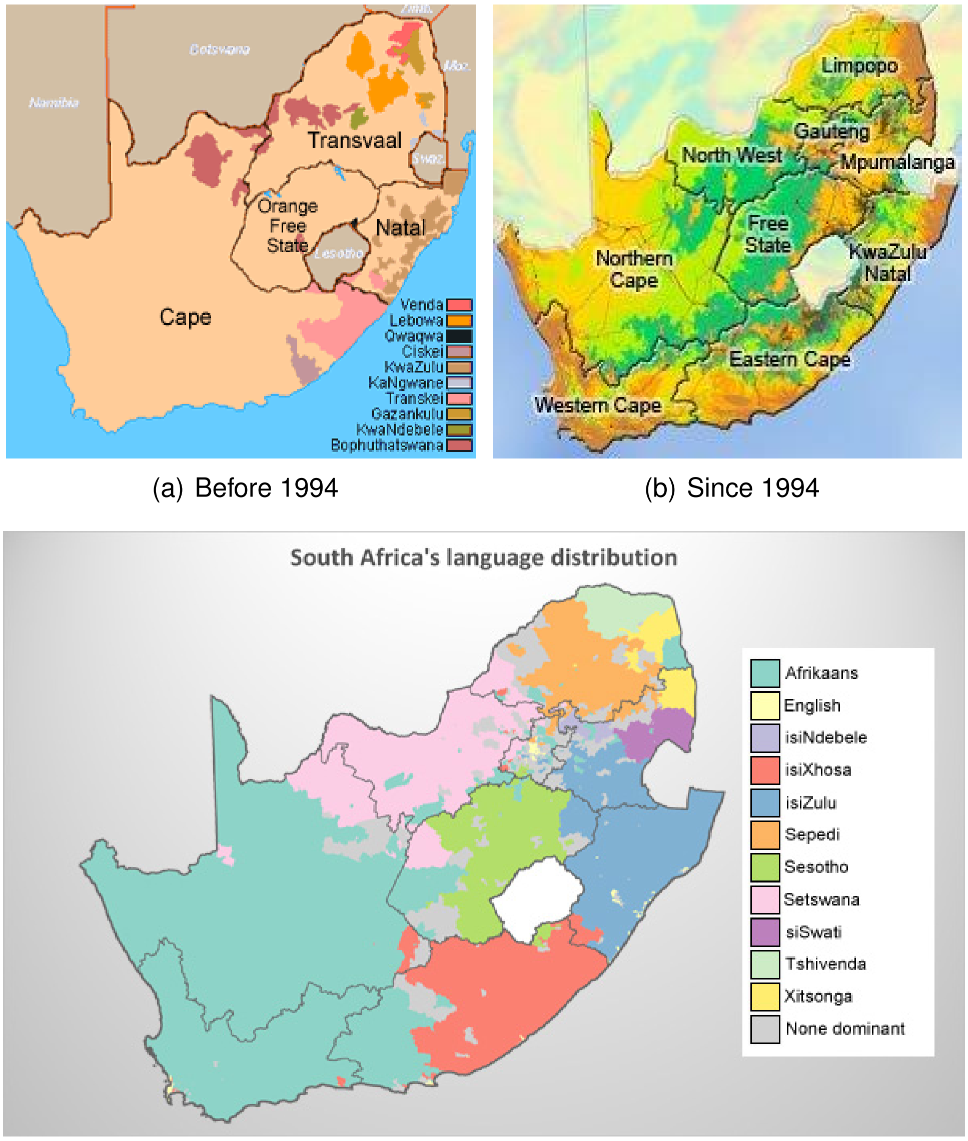
Chapter Preview

This chapter exposes the shortcomings of the use of descriptive statistics when analysing Grade 12 results. It also explains some of the factors, such as SES, parents' education level, and child and drug abuse, that affect learners' performance, and how learners' performances should be measured. The study's setting is briefly explained in terms of the South African setup. The definitions of hypotheses and the types of statistical errors are explained. A brief discuss regarding the testing of hypotheses in terms of point estimates and confidence intervals is outlined. The significance of this study is discussed and the chapter concludes by explaining how the thesis is arranged.

1.1 Introduction

The Republic of South Africa (RSA) has nine provinces, each with its own legislature, premier and executive council – and distinctive landscape, population, economy and climate. They are Eastern Cape, Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West and Western Cape. That is, under South Africa's new democratic constitution, the country was demarcated into nine provinces (see Figure 1.1).

Before 1994, South Africa had four provinces: the Transvaal and Orange Free State, previously Boer republics, and Natal and the Cape, once British colonies. Scattered about were also ”homelands”, some form of states to which black South Africans were forced to have citizenship. Figure 1.1 shows how South Africa has been demarcated into nine provinces from four pre-democratic provinces, and the distribution of the eleven official languages across the country.



(a) Before 1994

(b) Since 1994

(c) Distribution of Official Languages

Figure 1.1: Provinces of RSA before and since 1994, and language distribution
Source:*http://www.southafrica.info/about/geography/provinces.htm*

Each of South Africa's nine provinces has its own provincial government, with legislative power vested in a provincial legislature and executive power vested in a provincial premier and exercised together with the other members of a provincial executive council.

The provincial legislature has between 30 and 80 members elected for a five-year term based on the province's portion of the national voters' roll. The legislature is empowered to pass legislation within its functional areas.

The premier is elected by the legislature and, as with the President at national level, is limited to two five-year terms in office. The premier appoints the other members of the executive council (MECs). The executive council functions as a cabinet at provincial level. The members of the executive council are accountable individually and collectively to the legislature.

Presently, the responsibility for education is shared by two ministries, namely the Department of Basic Education (DBE) and the Department of Higher Education and Training, formed in 2009 when the former National Department of Education was split into two separate departments. That is, since 2009, the national Department of Education has been split into two ministries: Basic Education, and Higher Education and Training. The Department of Basic Education deals with all schools from Grade R to Grade 12[1], and adult literacy programmes, while the Department of Higher Education and Training deals with universities and other post-school education and training, as well as coordinating the Human Resource Development Strategy for South Africa.

1.2 Equal Education

In this era of globalisation and technological revolution, education is considered as a first step for every human activity; it plays a vital role in the development of human capital and is linked with an individual's well-being and opportunities for better living (Farooq, Chadhry, Shafiq and Berhanu, 2011; Kyei and Nemaorani, 2014). It ensures the acquisition of knowledge and skills that enable individuals to increase their productivity and improve their quality of life (Farooq et al., 2011). The quality of education and resources to support the education in South Africa prior 1994 were not distributed equally among citizens. In particular, majority of the citizens (Black South Africans) received inferior education.

It was acknowledged in the White-Paper (1995) that education like other commodities in RSA is not equally accessible to all citizens. The White-Paper (1995) states that ”for the first time in South Africa's history, a government has the mandate to plan the development of the education and training system for the benefit of the country as a whole and all its people. The challenge the government faces is to create a system that will fulfil the vision to *open the doors of learning and culture to all*. The paramount task is to build a just and equitable system which provides good quality education and training to learners, young and old, throughout the country”.

[1] In South Africa, Grade 12 is the final year of high school. It is more commonly referred to as matric, which is itself short for matriculation. At the end of Grade 12, students are said to be matriculated.

Kyei and Nemaorani (2014) lament that despite the attempts by the South African government to make education accessible to all by introducing free food and free textbooks, amongst others at the primary school level, the effects are not commensurate; for there is still a high failure rate (over 30%) and low retention rate, 44%, of high school learners in South Africa.

Ideally, it follows that the performances of learners in the provinces are expected to differ insignificantly if the distribution of resources is equitable; unless there is/are some other significant factor(s) that the authorities cannot account for.

1.3 Factors Affecting Learners' Academic Performance

The term 'academic performance' has been described as the scholastic standing of a learner at a given moment. It refers to how an individual is able to demonstrate his or her intellectual abilities. This scholastic standing could be explained as the grades obtained in a course or groups of courses taken (Adeyemi, 2008).

The performance of learners in national tests and national examinations is broadly used as an indicator of the effectiveness of the school. The results thereof have become such an acceptable indicator of school performance that for many, a school with high (weighted) examination or test scores is regarded as a good school (Barnard, 1999; Naidu, Joubert, Mestry, Mosoge and Ngcobo, 2008). However, there are many factors that contribute towards the learners' performance (Barnard, 1999; Karande and Kulkarni, 2005; Naidu et al., 2008).

For many years, schools and educators in South Africa have focused almost exclusively on learner performance. The educating community has been focusing on cognitive factors since more emphasis was paid to this aspect. Examples of cognitive factors are memory, verbal abilities and aptitudes for reasoning. These can be measured using performance and achievement tasks, where the answers given can be grouped as correct or incorrect (Johnson, 2000).

Among others, Johnson (2000) and Fan and Chen (2001) highlight that the educating community's focus is now on noncognitive factors because the realisation of these factors is evidenced. Here, we highlight some factors that affect learner performances.

1.3.1 Learner Background

Learner[2] background or family background plays a huge role in the process of learning. Family background includes factors such as socio-economic, parents' marital status (single, divorced or cohabit), family size, maternal characteristics, neighbourhood and parenting style (Marjoribanks, 1996). All these factors constitute the so-called home environment. It follows that home environment plays a crucial role in the learning process of individuals.

1.3.2 Impact of Divorce

Marital status of parents affects the performance of learners in school. In particular, divorce as one of several types of family problems, causes the distraction in the family structure - and this may cause the learner not to complete assignments or homeworks. Also, rumination about the divorce could cause lapses in a learner's concentration in the classroom (Schaffer and Schffer, 1997). Therefore, parents going through divorce may have children who experience increased likelihood for problems with social skills, behaviour issues, and academic achievement (Akanbi, 2014).

Children who are nine- to twelve-years old somewhat understand the divorce and are generally able to keep both their feelings and behaviour manageable. For this age group, anger is often the most powerful emotion. Children may physically act out their emotions and imitate family dynamics during play in order to cope with their feelings (Hughes, 2008; Akanbi, 2014). Smith (1999) reiterates that teenagers are generally a high-risk group during a family divorce. Teenagers are susceptible because they mourn as children; however, they are beginning to gain an understanding of the adult world and sometimes are conflicted in how they should show their emotions. In essense, learners' reaction to their parents' divorce varies based on the learner's age (Smith, 1999; Akanbi, 2014).

> *"When a baby is mirroring their parents' laughter or smiles, it is possible that babies can mimic similar sadness and anxiety when their parents feel those emotions also."*
>
> *Akanbi (2014, Page 106)*

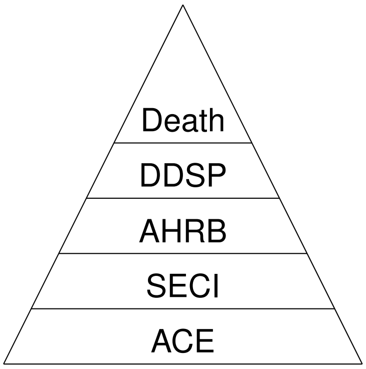
[2] The terms learner and student are used interchangeably. Likewise, educator and teacher.

1.3.3 Socio-Economic Status

Socio-economic status (SES) is mostly determined by combining the education level of parents, parents' occupational status and family income. SES has been found to be strongly associated with learner performances. Also, Gnanamoorthy (2014) found that there is a positive relationship between the SES of the parents (calculated in terms of family income, either by father, mother or both) and the academic achievements of their children. In other words, low-SES learners have fewer cognitive-enrichment opportunities (Jensen, 2009).

Figure 1.2, as given by Jensen (2009, Page 9) (read from the base to the vertex [bottom-up]), shows how adverse children experiences can set off an avalanche of negative life experiences, including social, emotional, and cognitive impairment; adoption of risky behaviours, disease, disability and social problems; and in worst cases, early death.

Figure 1.2: Adverse childhood experience model.
DDSP = Disease, disability and social problems,
AHRB = Adoption of health risk behaviour,
SECI= Social, emotional, and cognitive impairment,
ACE = Adverse childhood experience.



Common issues in low-income families include depression, chemical dependence, and hectic work schedules - all factors that interfere with healthy attachment that foster children's self-esteem, sense of mastery of their environment, and optimistic attitudes (Jensen, 2009).

Instead, poor children often feel isolated and unloved, feeling that kick off a downward spiral of unhappy life events, including academic performance (Jensen, 2009). In another words, low-SES learners score lower test scores and are more likely to drop out of school (Hochschild, 2003; Eamon, 2005). Arguably, SES dictates the quality of home life for children.

To be precise, poor learners bring many problems to school that more affluent learners usually avoid, all of which affect their readiness to learn and their ability to take advantage of what they are taught. These problems include poor health and nutrition, greater family instability, more frequent moves, less safe communities, fewer books and educational resources in the home or neighbourhood, a greater likelihood of having parents or other caretakers who have little formal education and/or speak little English, and anxieties about racial or ethnic discrimination (Hochschild, 2003).

Poor children are more likely to impose costs on society by consuming more health budget, more education resources, and more government economic aid. Because their chances of success are lower, they are also more likely to grow up to be poor themselves, thus perpetuating poverty into the next generation (Mayer, 2002).

1.3.4 Child and Drug Abuses, and Neglect

Child abuse refers to any recent act or failure to act on the part of a parent or caretaker which results in death, serious physical or emotional harm, sexual abuse or exploitation; or an act or failure to act which presents an imminent risk of serious harm (Christoffel, Scheidt, Agran, Kraus, McLoughlin and Paulson, 1992). In another words, child maltreatment refers to any non-accidental behaviour by parents, caregivers, other adults or older adolescents that is outside the norms of conduct and entails a substantial risk of causing physical or emotional harm to a child or young person. Such behaviours may be intentional or unintentional and can include acts of omission (i.e., neglect) and commission (i.e., abuse) (Christoffel et al., 1992; Corby, 2006). The terms child abuse and neglect, and child maltreatment are usually used interchangeably.

At birth, the brain is the most immature organ in the human body and will continue to develop as a result of nature or genetics and through environmental experiences. These events can have positive or negative consequences for healthy development (Terr, 1991).

Cognitive implications of child abuse include difficulties in learning and in school performance (Vondra, Barnett and Cicchetti,1990). Many studies have consistently stressed that abused, maltreated, or neglected children on the average score lower on cognitive measures and demonstrate lower school achievement when compared with their nonabused peers of similar socio-economic backgrounds (Vondra et al., 1990; Barnett, 1997).

The detrimental characteristics of abusive or neglectful parenting often lead to loss of self-esteem and a lack of motivation to succeed at school. At a very early age, maltreated children exhibit difficulties in self-esteem, behaviour, and adaptation to their environments. Abused toddlers respond more negatively, in contrast with nonabused peers, to their mirror images and make fewer positive statements about themselves (Barnett, 1997).

Substance abuse has negative effect on academic learning process (Schweinsburg, Brown and Tapert, 2008). In particular, marijuana's or dagga's negative effects on attention, memory, and learning can last for days or weeks after the acute effects of the drug wear off (Schweinsburg et al., 2008). Consequently, someone who smokes marijuana daily may be functioning at a reduced intellectual level most or all of the time. Not surprisingly, evidence suggests that, compared with their nonsmoking peers, students who smoke marijuana tend to get lower grades and are more likely to drop out of high school (Fergusson and Boden, 2008). Early drug or alcohol use may have potentially long-lasting consequences (Grant, 1998; Cleaver, Unell and Aldgate, 2011). Early onset of alcohol or other drug use is one of the strongest predictors of later alcohol dependence (Grant, 1998).

The effect of alcohol abuse is captured by Cleaver et al. (2011, Page 13) as follows:

> *"There is no doubt that alcohol misuse is associated with a wide range of problems, including physical health problems such as cancer and heart disease; offending behaviours, not least domestic violence; suicide and deliberate self-harm; child abuse and child neglect; mental health problems which co-exist with alcohol misuse; and social problems such as homelessness".*

1.3.5 Parenting

Parental effort is consistently associated with higher levels of achievement, and the magnitude of the effect of parental effort is substantial (Houtenville and Conway, 2008). Supportive and attentive parenting practices positively affect learners' academic achievement (Eamon, 2005). Parents from a more advantaged environment exert more effort, and this influences positively the educational attainment of their children (De Fraja, Oliveira and Zanchi, 2010). By the same token, the parents' background also increases the school's effort, which (in turn) increases the school achievement (De Fraja et al., 2010).

Family involvement is one of the most important contributors to school completion and success. The most accurate predictor of a learner's school achievement is the extent to which his/her family encourages learning. Success is more likely if the family communicates high, yet reasonable expectations for the student's education and future career and becomes involved in his/her education. Middle school and high school learners whose parents remain involved tend to (Clark, 1993; Henderson and Mapp, 2002; Mapp, 2004):

a) make better transitions,

b) maintain the quality of their work,

c) develop realistic plans for their future,

d) have higher graduation rates, and

e) advance to postsecondary education.

1.3.6 Parents' Education Level

Parents' education level plays a very important role in the performance of learners at school. Through multiple studies, the mother's educational level was a predictor of school completion for all middle adolescents participating in the studies (Halpern-Felscher, Connell, Spencer, Aber, Duncan, Clifford, Crichlow, Usinger, Cole, Allen and Seidman, 1997; Lacour and Tissington, 2011). In particular, the mother's education level has an effect on student academic achievement (Halpern-Felscher et al., 1997; Lacour and Tissington, 2011). In many studies, mother's education had a more significant effect on children's scores than income (Lacour and Tissington, 2011). In quantifying the effect of mothers' education level on learner achievement, Peters and Mullis (1997) found that the mother's education level had a 20% higher affect or effect than the father's education level on the academic outcomes of adolescents.

1.3.7 School Quintile

Learners in different classes and different schools do not have equal opportunities to learn: some learners may have unqualified educators, be in a school with limited resources, or possibly in poorly managed schools. These factors may contribute towards poor academic performance by learners (Stols, 2013). According to Stols (2013) quoting Kilpatrick, Swafford and Findell (2001); opportunity to learn (OTL[3]) is the most important predictor of learner achievement. Also, OTL remains the best explanation of the relationship between teaching and learning (Kilpatrick et al., 2001; Stols, 2013).

[3] The concept of OTL is used in different studies to determine or to quantify conditions within a school or classroom that promote or hamper learning. Some studies that measure OTL include factors like educator qualification, curriculum and materials, educators' professional development, safety and security of the learning environment, non-discriminatory policies, school financing, instructional practices, etc (Gillies and Quijada, 2008; Stols, 2013).

All South African public ordinary schools are categorised into five groups, called quintiles, largely for purposes of the allocation of financial resources. There are two steps in the classification of schools. First, a national poverty table, prepared by the Treasury, determines the poverty ranking of areas based on data from the national census including income levels, dependency ratios and literacy rates in the area.

Provinces then rank schools from quintile 1 to 5, according to the catchment area of the school. stricktly speaking, these poverty rankings are determined nationally according to the poverty of the community around the school, as well as, certain infrastructural factors. Schools in quintile 1, 2 and 3 have been declared no-fee schools, while schools in quintiles 4 and 5 are fee-paying schools.

The idea of free schooling is primarily about removing the financial barriers to education. Table 1.1 gives the National and Provincial breakdown of the quintiles. Each national quintile contains 20% of all learners, with quintile 1 representing the poorest 20% and quintile 5 the wealthiest 20% or the 'least poor'. However, provincial inequalities mean that these quintiles are unevenly distributed across provinces. The quintile ranking of a school is important because it determines the no-fee status of the school.

Table 1.1: The National and Provincial Breakdown of the Quintiles: National Poverty Table for 2014

Province	National Quintiles				
	One	Two	Three	Four	Five
Eastern Cape	27.3%	24.7%	19.6%	17.0%	11.4%
Free State	20.5%	20.9%	22.4%	20.8%	15.4%
Gauteng	**14.1%**	**14.7%**	**17.9%**	**21.9%**	**31.4 %**
KwaZulu-Natal	22.1%	23.2%	20.2%	18.7%	15.8%
Limpopo	28.2%	24.6%	24.2%	14.9%	8.0%
Mpumalanga	23.1%	24.1%	21.5%	17.7%	13.5%
Northen Cape	21.5%	19.3%	20.7%	21.4%	17.1%
North West	25.6%	22.3%	20.8%	17.6%	13.7%
Western Cape	**8.6%**	**13.3%**	**18.4%**	**28.0%**	**31.7%**
South Africa	**20.0%**	**20.0%**	**20.0%**	**20.0%**	**20.0%**

Source: http://wced.pgwc.gov.za/comms/press/2013/74_14oct.html

Table 1.1 shows that 8.6% of learners in the Western Cape Province fall into the category of learners in the poorest 20% in South Africa. It also explains why in the Western Cape Province, only just over 40% of schools are no-fee schools in quintiles 1, 2 and 3 when the average for South Africa as a whole is 60%. Similarly, 14.1% of learners in Gauteng Province fall into the category of learners in the poorest 20%

in South Africa. Also, only just over 46% of schools in Gauteng Province are no-fee schools in quintiles 1,2 and 3.

The quintile category of a school is a variable that is under the control of policy makers to alleviate the poverty status in schools, and therefore is a proxy for socioeconomic status or community characteristics. The policy context of this variable is viewed as the amount of money given to schools per learner, provision of nutrition programmes and non- payment of fees by parents.

Massive differentials on achievement tests and examinations reflect South Africa's divided past. Educational quality in historically black schools (quintiles 1 and 2) – constituting 80% of enrolment – has not improved since political transition, despite large resource transfers to such schools. Studies have found that school quintile is significantly associated with learner performance. For example, Van der Berg (2007) found that very poor schools (quintiles 1 and 2) were negative and significant in low performing schools and yielded different insignificant results elsewhere. Quintiles 3 to 5 were significant determinants of pass rates for low and middle performing schools, and were significantly strong to increase the dispersion on pass rates.

Also, Van der Berg (2007) showed that matriculation pass rates of schools were associated with pupil socio-economic background as measured by school fees, teaching resources (pupil/teacher ratio and average teacher salary), provincial location and the race category of schools. Socio-economic differentials play a major role in educational outcomes (at the primary school level) in South Africa (Van der Berg, 2008). More resources did not necessarily or without qualification improve school performance, although some resources (e.g. equipment at the school) appeared to play a role (Kurdziolek, 2011; Van der Berg, 2008). As in much of the educational production function literature, the message from Van der Berg (2008) study appeared to be not that resources did not matter, but rather that resources mattered only conditionally.

> *"Educators and policy makers believed that by providing more resources they could directly improve student-learning outcomes. To their frustration, this turns out not to be entirely true. Resources may be necessary but they are not sufficient. Resources themselves are not self-enacting, that is, they do not make change inevitable. Differences in their effects depend on differences in their use."*
>
> *(Kurdziolek, 2011)*

1.3.8 Class or School Attendance

The Department of Education, Taining and Employment (DETE) of Queensland undertook a study to establish the relationship between school attendance and learners' SES. Generally, many research outcomes confirm a strong link between attendance and learner SES. Compared to more affluent learners, children living in poverty are 25% more likely to miss three or more days of school per month (Ready, 2010).

While some learner absences are unavoidable and understandable due to illness and the like, or enforced through school disciplinary absences, many are not. These could be unexplained or unauthorised absences[4].

Poor school attendance can be linked to a number of related short term adverse outcomes and long term adverse outcomes for learners including lower academic outcomes, early school leaving, substance use, poverty, unemployment and negative health outcomes (DETE, 2013). However these factors may be interrelated in complex ways and factors that lead to low levels of attendance may also independently lead to some of these adverse outcomes.

A learner's regular absence from school may be a critical indicator in disengagement, leading directly to some of these adverse outcomes (DETE, 2013). Regardless of the nature of the relationship, poor school attendance, particularly with a high number of unexplained absences or unauthorised absences, is a readily observable warning sign for potential longer-term adverse outcomes.

The relationship exists between SES and school attendance, and this relationship is such that on average, learners from lower SES areas (neighbourhood) exhibit lower attendance rates (Ready, 2010; DETE, 2013). In particular, learners from lower SES backgrounds tend to exhibit higher levels of unauthorised and unexplained attendance (DETE, 2013).

There is a correlation between learners' attendance rates at school and academic performance. In particular, every day absent may be impacting on learner performance - thus, for school attendance, every day counts. However, this relationship does not subscribe to the notion of cause-and-effect, since the relationship is likely to be complex and impacted upon by the range of other factors (Balfanz and Byrnes, 2012; DETE, 2013).

[4] Classification of absence types: a) authorised - Illness, undertaking a medical procedure or attending a funeral, b) unauthorised - Bunking off classes, or c) unexplained - When no information has been provided by parents/carers or learners and the absence is pending the school's own investigations.

Learners miss school for a variety of reasons. By and large, however, these reasons fall into three broad categories (Balfanz and Byrnes, 2012):

1. Sometimes learners cannot come to school because circumstances or obligations compel them to be somewhere else during the school day.
2. On other occasions, learners will not attend school because they are actively avoiding interactions or events in school or on the way to or from school.
3. Finally, sometimes learners just do not go to school, not because there is something preventing or compelling them to stay away, but because they decide not to attend because they would prefer to be elsewhere, or just do not want to make the effort required to get to school.

1.3.9 Age and Gender

Demographic variables are thought to be having some effect on learner achievement. Studies have shown that as a learner gets older, the correlation between age and school achievement diminishes (Jabor, Machtmes, Kungu, Buntat and Nordin, 2011). The school provides an *equalising experiences* because the longer a learner stays in the school process, the more the effect of age on learner achievement is diminished (White, 1982; Jabor et al., 2011). This implies that the act of delaying school entry with the purpose of giving certain advantages to some learners is an exercise in self deception and it should be discouraged, for it could be a futile effort.

Okoro, Ekamen and Udoh (2012) highlighted that many studies have been conducted to investigate the effect of gender on learner performance. Gender of a learner plays a significant part in learner performance (Okoro et al., 2012). Awodun, Oni and Oyeniyi (2015: Page 73) lamented that a phenomenon in the school system that has been rather disturbing is the fact that despite the clamour for gender equality treatment, boys and girls do not seem to exhibit the same level of academic achievement. Depending on learning area for example, the overall science performance favours males (Jovanovic and King, 1998; Hedges and Newell, 1999; Demirbas and Demirkan, 2007; Nuzhat, Salem, Hamdan and Ashour, 2013). However, Hedges and Newell (1999) found that female learners had an advantage in reading and writing i.e., female learners had an advantage in fundamental learning areas (Jovanovic and King, 1998; Demirbas and Demirkan, 2007; Nuzhat et al., 2013; Awodun et al., 2015). The issue of gender comparisons should be treated with causion. Demirbas and Demirkan (2007) and Sunday and Zaku (2013) warn that findings on gender are inconclusive since different researchers have different opinions based on their findings regarding gender effect on learners' academic performance.

The effects of the educator and the gender of educator cannot be ignored. Bansilal, James and Naidoo (2010) have shown that the effect of an educator can be negative or positive depending on the space and time it happens. Example of the effect of educator on the learning process of learners (Bansilal et al., 2010):

> *"Most of the Grade* 11 *mathematics learners in the study conducted by Moodley (2008) indicated that their mathematics teachers displayed a negative attitude towards them. For example,* 91% *(29/32) of her sample indicated that the teacher ignored them when they asked questions and* 93% *(30/32) indicated that the teacher made them feel silly when they asked questions in the maths class.*
>
> *Some of the learners' comments include: "He tries to be funny but he doesn't know that he actually embarrasses and hurts people" and "You know you afraid to ask questions. Maybe the teacher will make you feel stupid" and further, "I hate being looked down upon". These comments support the contention made by the learners in our study that some teachers sometimes make negative comments to learners and when this happens, they are embarrassed and feel belittled by these disparaging comments."*

Okoro et al. (2012) assert that with the presence of male educator diminishing globally it was important to investigate the effect of male educator on the performance of learners in Nigeria. Okoro et al. (2012) investigated the following questions:

a) To what extent do male learners taught by male educator differ from those taught by female educators in their academic performance?

b) To what extent do female learners taught by male educators differ from those taught by female educators in their academic performance?

c) To what extent do learners taught by male educators differ from those taught by female educators in their academic performance?

Okoro and Uwah (2013) found that educator gender has a significant influence on the academic performance of learners. Specifically, male learners taught by male educators perform significantly differently from male learners taught by female educators. Similarly, boys taught by male educators have a more positive attitude to schooling than boys taught by female educators (Okoro and Uwah, 2013), which (in a way) may translate into better learner performance.

Disparity also exists in the performance of female learners taught by male educator and those taught by female educator (Okoro et al., 2012).

1.4 Measuring Learners' Academic Performance

Student learning performance is influenced by, among other things, the facilities provided, see Figure 1.3. Learning performance not only focuses on learner results, but it is also related to the other contributions. Therefore, the terms of learning performance have a very broad definition. Table 1.2 gives elements that are related to the terms of learning performance (Castaldi, 1982; Mendell and Heath, 2005).

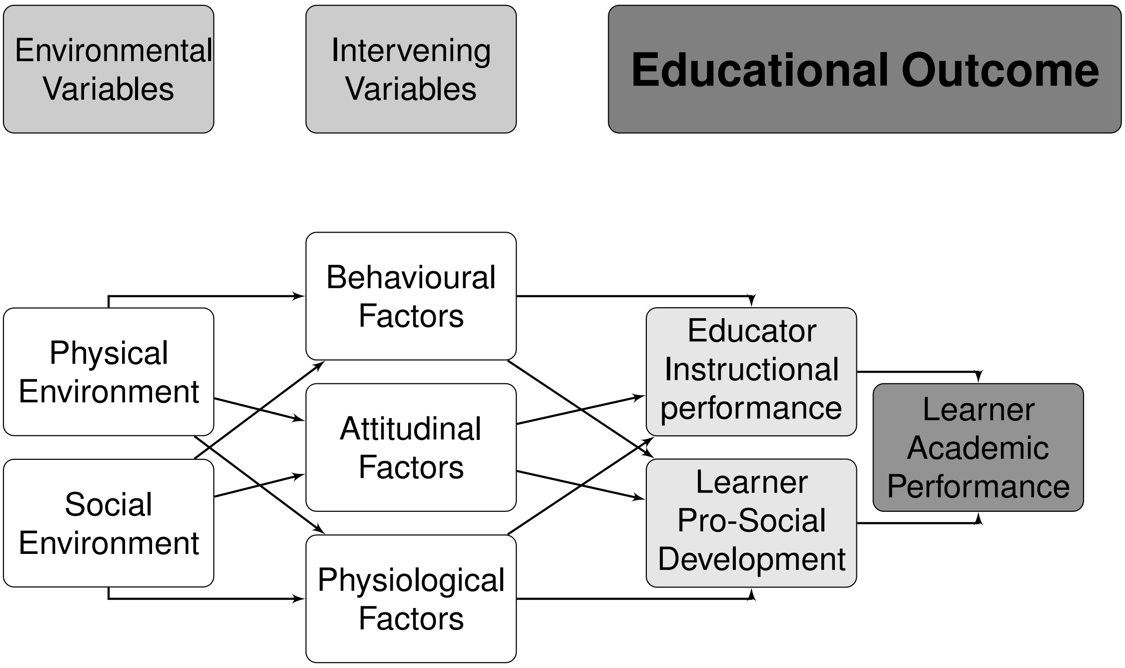


Figure 1.3: Links between Environmental and Educational Outcomes.
Source: Lackney (1999)

Table 1.2: Elements related to the terms of learning performance

Social Development	Individual Differences and Group Similarities
Multi-stimuli Instruction	Attending and Learning
High Transfer of Learning	Nurturing Readiness
Promoting Motivation	Anomalous Students
Reducing Fatigue and Improving Learning	Promoting Activity Programmes and Meaningful Learning
Effective Group Instruction	Incidental Learning

The elements stated in Table 1.2 have contributed to the students' learning performance. Hence, it is important to provide effective facilities to the students for their learning activities, in order to increase their performance.

Evaluating what learners have learned throughout the course can be accomplished in many ways, depending on the course objectives and how learner performance will be measured. Measuring performance and achievement is a significant part of the education process and informs educators of learner ability and progress toward educational goals. It is also the primary gauge used by educators to guide the advancement of learners through the education process.

Methods for measuring learner performances are diverse (Mushtaq and Khan, 2012). Formal tests, quizzes, and examinations are the traditional methods for assessing learner achievement. The choice of method depends on the focus of the study or investigation (Mushtaq and Khan, 2012). The most common method used by researchers is the use of test results or previous year results since they are studying performance for the specific subject or year (Hijazi and Naqvi, 2006). Similarly, educational services are often not tangible and are difficult to measure because they result in the form of transformation of knowledge, life skills and behaviour modifications of learners (Farooq, Chaudhry, Shafiq and Berhanu (2011). So there is no commonly agreed upon definition of quality that is applied to education field since it varies from culture to culture (Farooq et al., 2011, Page 2).

> *“Today’s school-reform initiatives often center on using measures of student learning to gauge school and educator effectiveness. This focus on accountability has in some ways been taken away from the more basic purpose of assessment: to figure out what students or learners know and need to learn”*
>
> *http://blogs.edweek.org/teachers/teaching_ahead/what-are-the-best-ways-to-measure-student-learning/*

Learner performance is influenced by many factors. Broadly, these factors may be categorised as environmental factors and intervening factors (Lackney, 1999; Farooq et al., 2011). Intervening factors are mostly embodied by a learner while environmental factors are those that a learner happened to find himself/herself in e.g., social and physical contexts. In essence, as highlighted by Lackney (1999), intervening variables are mostly and directly affected by environmental variables while educational outcomes (learner performances) are directly affected by intervening variables. Figure 1.3 attempts to simplify this explanation and summarises Section 1.3 and Section 1.4.

Assessment and evaluation are essential components of teaching and learning. Without an effective evaluation programme it is impossible to know whether students have learned, whether teaching has been effective, or how best to address student learning needs. The quality of the assessment and evaluation in the educational process has a profound and well-established link to student performance. The terms ‘Evaluation’ or ‘Assessment’ are both in common usage in educational circles and sometimes are being used interchangeably. Taras (2005) clarifies the process of assessment.

Assessment or evaluation is of central importance in education (Carr, McGee, Jones, Mckinley, Bell, Barr and Simpson, 2004). Assessment is the process of gathering information on student learning. Evaluation is the process of analysing, reflecting upon and summarising assessment information, and making judgements and/or decisions based on the information collected (Taras, 2005). Reporting involves communicating the summary and interpretation of information about student learning to various audiences who require it. Bansilal et al. (2010, Page 155) quoting Carr et al. (2004) emphasise the purposes of assessment as follows:

> *"Assessment is an integral part of the learning process and has both formative and summative functions. These two sets of functions are mainly a matter of when they occur in relation to their purpose, and not a differentiation of rigour or quality. Formative assessment is an on-going informed interaction between the teacher and student designed to enhance student learning. Therefore it provides feedback to the teacher and to the student about present understanding and skill development in order to determine a way forward."*

1.4.1 Grade 12 Results

In South Africa, the National Senior Certificate (NSC) examinations are written after twelve years of formal schooling and signify the end of the Further Education and Training (FET) band. Grade 12 results are used to measure the competency of learners and the effectiveness of schools. However, the authorities rely on results from descriptive analysis to make inference.

The Grade 12 pass rate.

Grade 12 results should be treated with caution, even though the pass rate[5] can serve as a measure of opportunities open to the youths. It is noted that the "Grade 12 pass rate on its own is not a good measure of academic achievement in the schooling system, nor was the pass rate ever designed for this", concedes the Department of Basic Education (DBE) of South Africa. However, the pass rate can serve as a measure of the opportunities open to our youths. If these opportunities increase, then we should celebrate.

[5] Pass rate refers to the proportion (converted to percentage) of learners who have passed Grade 12 in a particular academic year.

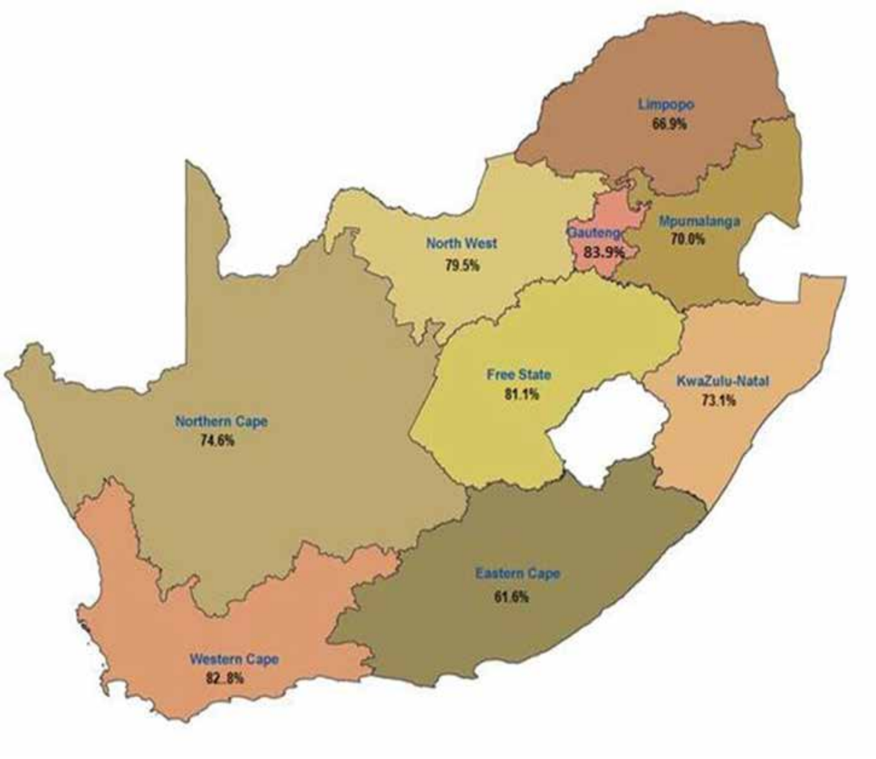


Figure 1.4: The 2012 Provincial Pass Rates in RSA
Source: Department of Basic Education

> *"Matric results, which have been getting better for several years, are still improving. The national pass rate in* 2012 *rose to* 73.9%*, up by more than* 13% *points from the* 60.6% *who achieved a pass in* 2009 *and by more than three percentage points from the* 70.2% *who did so in* 2011*".*
>
> *https://africacheck.org/reports/minister-wrong-to-say-better-results-show-sa-education-is-improving/*

Clearly, from the extract given above it cannot be seen how confounding was accounted for, and reliance only on descriptive statistics is not encouraged. Figure 1.4 presents summary statistics on 2012 pass rates, according to provinces.

It can be seen from Figure 1.4 that Gauteng Province and Western Cape Province performed marginally higher than the rest of the provinces. These two provinces have consistently been performing marginally higher or better than all other provinces over the years, hence this study focused on them (Gauteng Province and Western Cape Province).

To celebrate the marginal differences between the provinces is an exercise in self-deception, especially when the results are only based on descriptive statistics. The only time we should celebrate is when we acknowledge the methods used to analyse the data, and justify in our methodology and statistical approaches that the analysis accounts for interrelatedness of our observations. Also, if the chosen analytical technique account for confounding effects. Chapter 4 demostrates the statistical approaches for dealing with clustered data (such as Grade 12 results).

1.5 Research Approach and Design

This is a special study in that it is mainly a methodological study. It proposes an approach towards the analysis of Grade 12 results in the South African Education system. This orientation makes it a challenge to follow a traditional design in which issues pertaining to literature review, research methodology and research findings are distinct and clearly demarcated into separate chapters. In the current case, an integrated approach was found more appropriate.

This study followed a quantitative approach, the one in which the investigator primarily uses post-positivist claims for developing knowledge (i.e., cause and effect thinking, reduction to specific variables and hypotheses and questions, use of measurement and observation, and the testing of theories), employs strategies of inquiry such as experiments and surveys, and collects data on predetermined instruments that yield statistical data (Creswell, 2003). In particular, population-averaged models, which are computer intensive, were employed to compare the performances of Gauteng Province and Western Cape Province.

A combination of computer softwares packages, Excel (data management) and Stata (data analysis), was used to accomplish this task.

This research approach was implemented on the secondary data, Grade 12 data as supplied by Umalusi[6]. Therefore, no data collection instrument, such as a validated questionnaire or data faxing, was used to collect the data.

1.6 Aim of the Study

Grade 12 examination is regarded as the crucial step as it is used as a yard stick for deciding whether or not a learner is ready or prepared for tertiary education. It is the Grade 12 results that generate more interest to the public such that at the minute (micro) level - schools are compared, at local level - regions are compared, and at

[6] Umalusi is the quality assurer in the general and further education and training bands of the national qualifications framework (NQF). The Council ensures that the providers of education and training have the capacity to deliver and assess qualifications and learning programmes and are doing so to expected standards of quality. Umalusi is guided by the General and Further Education and Training Act, Act 58 of 2001, published in December 2001.

(macro or) national level - provinces are compared in terms of how their respective learners performed.

Therefore, there is a need to have a reliable method(s) for comparing the performances of provinces given that the intervention (which is the teaching of learners) was not targeted, and that the setup was not controlled. The possibility of spill-over of the intervention effect between provinces is highly likely. Thus, the groupings are not randomised, and controlled.

The study is aimed at proposing the method that will plausibly be used to analyse the Grade 12 results, and (the method used) to compare performances of learners between provinces of the Republic of South Africa.

The researcher intends to employ causal models for comparing performances of nonrandomised study groups in South Africa. Such analysis was never performed as the South African government, like other states in Africa e.g., Tanzania (Kassile, 2014), Zimbabwe (Nyagura, 1991) and Botswana (BEC, 2011), and other states outside Africa e.g., Singapore (MES, 2014) and United Kingdom (DFE, 2010) relied on descriptive analyses. Therefore, these studies used descriptive statistics (proportions expressed as percentages) as a measure of school performance, e.g., see Figure 1.4, which most of the time the analyses ignored the intracluster correlation coefficient (ICC), a measure of the relatedness of clustered data. It accounts for the relatedness of clustered data by comparing the variance within clusters with the variance between clusters.

It is in the public domain that education is one of the present South African government's priorities - the distribution of resources e.g., quality educators (high effective educators) and infrastructure (well-funded instructional materials), among schools will be in such a way that it is properly distributed (equitable); for this study will have provided the authorities with supporting evidence. Therefore, the issue of school quintile will be revisited, probably this concept may be phased-off in the near future.

In this study we (attempt to) address the following research questions:

- What insights are revealed through the application of causal models in analysing the effect of gender and school quintile in the performance of learners?

- How do we apply the population average model in case of nonrandomisation studies in educational setting?

1.7 Hypothesis

In this study we (attempt to) test the hypothesis that the performances of learners in the two provinces, Gauteng and Western Cape, are not significantly different, and the interpretation is performed at 95% confidence limit (2-sided).

A hypothesis[7] is a tentative statement about the relationship between two or more variables. In another words, a hypothesis is a specific conjecture (statement) about a property of a population of interest. A hypothesis is a specific, testable prediction about what one expects to happen in one's study. While the hypothesis predicts what the researchers expect to see, the goal of research is to determine whether this guess is right or wrong.

When trying to come up with a good hypothesis for one's own research or study, one needs to take heed of the following questions:

- Is one's hypothesis based on one's research topic?
- Can the stated hypothesis be tested?
- Does the hypothesis include independent and dependent variables?

Clearly, hypothesis is highly related to literature review, study design and the nature of the collected data and the methods to be used for data analysis.

1.7.1 The Null and Alternative Hypotheses

In designing a study it is important to have a clear research question and to know the outcome variable to be compared. Once the research question has been stated, the null and alternative hypotheses can be formulated (Julious, 2010). In another words, in order to determine if the results of the study are significant, it is essential to also have a null hypothesis (Larson, 1982; Julious, 2010). The null hypothesis (H_0) is the prediction that one variable will have no association to the other variable. In other words, the null hypothesis is usually of the form of no difference in the outcome of interest between the study groups (Larson, 1982). The study or alternative hypothesis, (H_1), would then usually state that there is a difference between the study groups (Julious, 2010).

[7] **Definition**: A simple hypothesis H is any statement that completely specifies the probability law for a random variable X. A hypothesis that is not simple is called composite. A test of hypothesis, H, is any rule that tells us whether to accept H or reject H, for every possible observed random sample of X (Larson, 1982).

The null hypothesis is assumed to be valid unless contradicted by the results. The experimenters can either reject the null hypothesis in favour of the alternative hypothesis or not reject[8] the null hypothesis. To say that you are accepting the null hypothesis is to suggest that something is true simply because you did not find any evidence against it. This represents a *logical fallacy*[9] that should be avoided in scientific research (Gunderman and Sistrom, 2006).

1.7.2 Type I and Type II Errors

Two types of errors occur when testing hypotheses. If the null hypothesis is rejected when it is true, then a Type I error has occured. If the null hypothesis is not rejected when it is false, then a Type II error has been made. The probabilities of Type I error and Type II error are called alpha (α) and beta (β), respectively (Altman, 1991, Page 169; Zhou, Obuchowski and McClish, 2011, Page 21).

The probabilities of making Type I and Type II errors are given as, for example (Larson, 1982, Page 414; Chow and Liu, 2000, Page 127; Zhou et al., 2011, Page 22) :

$$\begin{aligned} \alpha &= P(\text{Type I error}) \\ &= P(\text{reject } H_0|H_0 \text{ true}) \end{aligned} \tag{1.7.1}$$

$$\begin{aligned} \beta &= P(\text{Type II error}) \\ &= P(\text{accept } H_0|H_1 \text{ true}) \end{aligned} \tag{1.7.2}$$

The probability of making Type I error, α, is called the level of significance. Table 1.3 , as given by Julious (2010, Page 11), summarises the relationship between the two types of errors. *Power* of the statistical test is the probability of correctly rejecting the null hypothesis H_0 when H_0 is false; that is,

[8] It is important to remember that not rejecting the null hypothesis does not mean that you are accepting the null hypothesis.

[9] There are many different types of logical fallacy including:

- Formal Fallacy: These are also called deductive fallacies. In deductive fallacy arguments, all premises must be accurate and impossible to be proven otherwise. When this is the case, there is no way that the conclusion can be false.
- Informal Fallacy: This is an inductive argument.
- Logical and Factual Errors: Any argument in which premises or inferences are poor will result in a fallacious conclusion.

$$\begin{aligned} Power &= 1 - \beta \\ &= P(\text{reject } H_0 | H_0 \text{ false}) \end{aligned} \tag{1.7.3}$$

On the other hand, the power of a study reflects the probability that the study will be able to detect a true effect of a specified size. So, given that an effect exists, the power of the study is the likelihood that an effect will be found in the study's results. It is the probability of rejecting the null hypothesis when it is in fact false.

We note that statistical power is affected by three factors (Suresh and Chandrashekara, 2012):

i) The difference in outcome rates between the two groups. A smaller difference requires exponentially more power.

ii) The level of significant difference one hopes to show (e.g. $p < 0.05$ or < 0.001). Chasing after a small p value takes more study power.

iii) The frequency of the outcome in the two groups. Imagine an exposure that increases incidence by a third: it is easier to show a difference between 30 and 45 percent than between 10 and 15 per cent. Maximum power is reached when roughly half of the people studied have the outcome of interest.

Table 1.3: Relationship between the H_0 and H_1

	If H_0 is	
	True	False
Fail to reject	No error	Type II error
Reject	Type I error	No error

The following criteria are commonly used as a rule of thumb for choosing the null hypothesis (Chow and Liu, 2000, Page 129):

(a) Choose H_0 based on the importance of Type I error. Under this rule, it is believed that a Type I error is more important and serious than Type II error. We would like to control the change of making a Type I error (i.e., α) at a tolerable limit. Thus, H_0 is chosen so that the maximum probability of making a Type I error [i.e., P(reject H_0 when H_0 is true)], will not exceed α level.

(b) Choose the hypothesis we wish to reject as H_0. The aim, here, is to establish the alternative hypothesis H_1 by rejecting H_0.

1.7.3 Hypothesis Testing: The P-value

A P-value[10] is the probability of obtaining the study results (or results more extreme) if the null hypothesis is true (Altman, 1991, Page 167). In practice, what happens in a research study is that the null hypothesis of no difference between the two study arms is stated, that is, $\mu_A = \mu_B$. The study is then conducted, and a particular difference Δ is observed such that $\Delta = \hat{\mu}_A - \hat{\mu}_B$. Here, μ_A and μ_B are averages in study arms A and B, respectively, and $\hat{\mu}_A$ and $\hat{\mu}_B$ are mean estimates in study arms A and B, respectively.

Due to pure randomness even if the two study arms are truly the same we would seldom actually observe $\Delta = 0$ but some random difference (Julious, 2010). If Δ is small then the probability of seeing this difference under the null hypothesis could be very high. If a larger difference is observed, then the probability of seeing this difference by chance is reduced. As the difference increases, the P-value decreases (Julious, 2010).

A small P-value indicates that the results obtained are unlikely when the null hypothesis is true, and the null hypothesis is rejected. Conventionally the cut-off value or two-sided significant is set at 0.05 or 5% (Altman, 1991). Table 1.4, as given by Julious (2010, Page 9), attempts to summarise this information.

Table 1.4: Statistical significance

	P-value < 0.05	**P-Value** ≥ 0.05
Result is	Statistically significant	Not statistically significant
Decision	Sufficient evidence to reject the null hypothesis	Insufficient evidence to reject the null hypothesis

In summary:

> Hypothesis testing is a scientific process to examine whether or not a hypothesis is plausible. In general, hypothesis testing follows the next five steps.
>
> 1) State the null and alternative hypotheses clearly (one-tailed or two-tailed test).
> 2) Determine a test size (significance level). Pay attention to whether a test is one-tailed or two-tailed to get the right critical value and rejection region.

[10] Since the P-value is the probability - its values vary between 0 and 1. That is, $0 \leq$ P-value ≤ 1.

3) Compute a test statistic and P-value or construct the confidence interval, depending on testing approach.
4) Decision-making: reject or do not reject the null hypothesis by comparing the subjective criterion in 2) and the objective test statistic or p-value calculated in 3).
5) Draw a conclusion and interpret substantively.

There are three approaches to hypothesis testing: namely Test statistic approach, P-value approach and Confidence interval approach (as presented by Table 1.5). Each approach requires different subjective criteria and objective statistics but ends up with the same conclusion (Larson, 1982).

Table 1.5: Approaches of hypothesis testing

Step	Test Statistic Approach	P-Value Approach	Confidence Interval Approach
1	State H_0 and H_1	State H_0 and H_1	State H_0 and H_1
2	Determine test size α and find the critical value.	Determine test size α.	Determine test size α or $1-\alpha$, and the hypothesised value.
3	Compute a test statistic	Compute a test statistic.	Construct the $(1-\alpha)*$ 100% confidence interval.
4	Reject H_0 if $TS^{a} > CV^{b}$	Reject H_0 if p-value $< \alpha$.	Reject H_0 if a hypothesised value does not exist in confidence interval.
5	Interpretation	Interpretation	Interpretation

[a] *TS = Test statistic*
[b] *CV = Critical value*

The hypotheses in this study are:

a) A case of *averages*:

- H_0: The average performances of learners in the two provinces are not statistically different ($\mu_1 = \mu_2$).
- H_1: The average performances of learners in the two provinces are statistically different ($\mu_1 \neq \mu_2$).

b) A case of *proportions*:

- H_0: The proportion of learners who passed Grade 12 to the proportion of learners who failed Grade 12 in the two provinces are not statistically different ($p_1 = p_2$).

- H_1: The proportion of learners who passed Grade 12 to the proportion of learners who failed Grade 12 in the two provinces are statistically different $(p_1 \neq p_2)$.

We note that $\mu_{i\in\{1,2\}}$ and $p_{i\in\{1,2\}}$ are the the true mean and true proportion, respectively.

1.8 The Significance of the Study

An informed decision requires relevant and useful information. For example, decisions about major consumer purchases are often preceded by reviews of Consumer Reports or similar compilations of ratings or measures of features and characteristics of the product to be purchased (Brown, Wohlstetter and Liu, 2008). The issue of school choice or province choice, with regard to learner performances, is no different. Decisions about choosing which province is most appropriate, best performing, or more desirable within a given community also require a collection of relevant and useful information that best reflects the important features and characteristics of the province (Brown et al., 2008).

A lot of research has been done on factors affecting academic performance of learners (basic education) and college/university students; for example Mushtaq and Khan (2012). However there is scarce information about proper comparisons of performances of learners according to their respective strata. In the South African setting - the strata will refer to provinces, regions, and schools. The study will enable the researcher to make recommendations to South African DBE policy makers especially those that deal with issues regarding quality assurance (QA), and the Ministry of Education on what policies and strategies can be employed to improve academic performances of learners in the provinces.

This study should be particularly significant, as it is the first cross-sectional study to be conducted that employs causal models to compare performances of provinces in terms of Grade 12 results in RSA. The importance of conducting this research is to demonstrate the method that is suitable for comparisons of nonrandomised groups. In particular, the researcher will demonstrate the use of causal models in the analysis of clustered data such that:

(i) The educational authorities, e.g. curriculum advisors, circuit managers and/or monitoring support groups draw a better conclusion with regard to Grade 12 performance.

(ii) Concerned authorities will understand and appreciate these approaches and possibly come up with the intervention strategies to be implemented in improving the academic performance at their schools, regions and provinces.

(iii) It will also be a contribution to the body of knowledge in the field of application of causal models in the evaluation of interventions in nonrandomised groups or social science settings.

(iv) This study will also serve as a basis for other related research that may affect academic performance in centers of learning i.e., schools, and tertiary institutions.

In another words, currently the approach to the analysis of Grade 12 results in South Africa is limited to the use of descriptive statistics. The purpose of this study is to use the potential power of causal models (in particular, population average models) to better explain Grade 12 results. Consequently, this will provide a better platform upon which strategic interventions can be developed. The study also seeks to demonstrate how inference is drawn in case of nonrandomisation studies in educational setting.

1.9 Ethical Consideration

Research ethics refers to the application of moral standards to decisions made in planning, conducting, and reporting the results of research results (McNabb, 2010, Page 69). Research ethics provides guidelines for the responsible conduct of (biomedical) research (Welman, Kruger and Mitchell; 2008). In addition, research ethics educates and monitors scientists conducting research to ensure a high ethical standard. Ethical considerations, which range from plagiarism, respect for human rights when data is collected to honesty in reporting research results, have become an intergral element of every research proposal and study (Welman et al., 2008; Drake and Heath, 2011, Page 47).

Plagiarism is the use of others' data or sources or ideas without due acknowledgement or permission (Welman et al., 2008, Page 182).Therefore, in this study all sources used were duly acknowledged.

As highlighted by Welman et al. (2008, Page 181); there are three stages that should be observed when dealing with respect for human rights during the data collection process; namely:

- when participants are recruited,

- during the intervention and/or measurement procedure to which they are subjected, and
- in the release of results obtained.

In this project secondary data was used. Therefore, no data collection tool such as questionnaire was used and as such no participants were recruited.

McMillan and Schumacher (1993) maintained that it is imperative for researchers to obtain permission to enter any particular field and ensure the confidentiality and anonymity of the participants, thus encouraging the latter's free choice of participation. This required a full description and disclosure to the participants of how the data that was collected was intended to be used by the researcher. The confidentiality and anonymity of participants or learners was guaranteed by not revealing their names, student numbers, examination numbers or any data against their will. In particular, psuedo-identifiers were used to blind the researcher. The statistical results were reported objectively.

The ethics clearance (Project No.: TREC/135/2015: PG) for this project was granted by Turfloop Research Ethics Committee (TREC).

1.10 Arrangements of the Chapters in the Thesis

This thesis is sub-divided into 5 chapters with Chapter 1 providing the background to the setting wherein the study will be focused. This chapter outlines how South Africa and its population is distributed, and as such the issue of comparisons of learner performances come into effect. Factors that affect the performances of learners are discussed, the research design and research approach to this study is discussed. Some pointers with regard to the significance of this study are outlined and how the hypothesis are set in this study is entertained in details.

Chapter 2 deals with challenges that one faces when dealing with the concept of causality in observational studies. In particular, the description of problem of causal inference is discussed, and the Rubin causal model is introduced. The fundamental of causal problem is discussed together with assignment mechanism. The discussion of causal effect in both randomised and nonrandomised studies is then outlined. The importance of matching and how matching methods are implemented, and the introduction of propensity score and how it (propensity score) is estimated are introduced.

Chapter 3 introduces the concept of clustered data, and how statistical models are implemented in analysing such data. Ordinary regression models, linear and logistic regression models, are introduced and how the point estimates from these models are determined and interpreted. Mixed models and hierarchical models are introduced and discussed in this chapter. How clustered data are statistically modeled using linear mixed-effects models and nonlinear mixed-effects models is outlined; as the theoretical background to these models is presented. The challenges posed by spill-over effects or contamination are discussed since social science experiments are not controlled experiments.

Chapter 4 presents data analysis and interpretation of the results. The results are based on descriptive analysis and application of crude models and adjusted models. The results are presented in tabular and graphical formats, and Chapter 5 presents the conclusion and recommendations.

Chapter Summary

The need for review of the approaches to analysis of Grade 12 results was described; for in RSA - Grade 12 results are used to measure the competency of learners and the effectiveness of regions and schools.

A discussion of factors that affect learners' performances was detailed and learned how educational outcomes (educator instructional performance and learner pro-social development) are related to intervening variables (behavioural factors, attitudinal factors and physiological factors) and environmental variables (physical environment and social environment).

The justification for the choice of quantitative study design has been given, and in particular the Ex Post Facto design, as adopted in this thesis. We have become acquainted with statistical hypothesis testing, Types I and II errors, and power of study. Finally, the definition of causal model was given: a statistical tool for evaluating the effect of 'intervention'. A precise outline for the need to apply causal models in the analysis of clustered data such as Grade 12 results was supplied.

Chapter 2

Problem of Causality in Observational Studies

*****

"We may define a cause to be an object, followed by another, ...
where, if the first object had not been, the second had never existed."
Hume (1748)

Chapter Preview

Chapter 2 discusses causality in observational studies. The problem of causal inference is highlighted given the baseline covariates in the study arms. Then the challenges that researchers face when dealing with causality in observational studies are highlighted. The potential-outcomes framework for causal inference together with assignment mechanism are discussed. The importance of matching and how matching methods are implemented are detailed. Finally, the technique of propensity score and how propensity score is estimated are discussed.

2.1 Introduction

An observational study, a study that draws inferences about the possible effect of a treatment on subjects where the assignment of subjects into a treated group versus a control group is outside the control of the investigator, entails a crucial handicap for estimating causal treatment effects in comparison to a randomised controlled trial (RCT). Obviously, observational study lacks the random assignment of treatment and therefore it is often referred to as a nonexperimental or a quasi-experiment (Rubin, 1974).

The random treatment assignment assures that, in expectation, the distributions of covariates are similar between treated and untreated individuals such that they are comparable (Shadish, Cook and Campbell, 2002; Rosenbaum, 2005). In another words, one of the key benefits of randomised experiments for estimating causal effects is that the study groups are guaranteed to be only randomly different from one another on all background covariates, both observed and unobserved (Shadish et al., 2002; Rosenbaum, 2005; Stuart, 2010). Differences in effect estimates are hence completely attributable to the treatment. It can be expected that both covariates measured at baseline, i.e., measured before treatment assignment, and those remaining unmeasured are all balanced between treatment groups such that overt as well as hidden bias can be avoided in RCTs. Due to numerous reasons, e.g., financial, ethical or temporal, RCTs are not always feasible or they are not the appropriate study type such that observational studies are used to investigate causal treatment effects (Rosenbaum, 2005).

Since the treatment assignment is not (at) random but depends on participants‘ covariates, observational studies always require an adjustment for imbalances in covariate distributions between treatment groups. This is often done, e.g., by using traditional regression techniques or matching by relevant covariates (Rubin, 1979; Robins, 2005). However, this is not free of trouble when there is a large set of covariates for which it has to be adjusted for and for which matching should be done (Rosenbaum, 1989).

Also, only overt bias, i.e., bias induced by measured covariates, can be accounted for in observational studies independent of the method preferred for imbalance adjustments. Hidden bias, i.e., bias caused by covariates remaining unmeasured, cannot be controlled such that additional assumptions need to be met to still obtain reliable causal treatment effect estimates in observational studies (Rosenbaum, 1991). Additionally, the distinction between causal and associational parameters and their way of estimation have to be kept in mind.

In consequence of a discussion about the validity of observational studies regarding the estimation of causal effects in comparison to RCTs, Rubin (1974) investigated both study types comparatively from the causal point of view. He concluded that both observational studies and RCTs are suitable to estimate causal effects, but he also highlighted that one study type’s advantage is another study type’s drawback: that there is no clear answer to the question which study type should be preferred to estimate causal effects (Rubin, 1974).

The generalisation of results from observational studies is more obvious due to its non-experimental design compared to RCTs. D’Agostino (2007) agreed and stated that RCTs provide rather efficacy (treatment effects estimated in an RCT setting, e.g., restricted by inclusion and exclusion criteria) than effectiveness (effect of treatment which would be measured in circumstances more similar to a real world setting) due to its limited circumstances. However, the apparent benefit of observational studies, i.e., the generalisability of their results, can be substantially reduced if there is only an erroneous or even missing control for bias, especially for hidden bias.

The discussion about the ability of observational studies to estimate reliable causal effects has been ongoing (McKee, Britton, Black, McPherson, Sanderson and Bain, 1999; Pocock and Elbourne, 2000; Ioannidis, Pantazis, Kokori, Tektonidou, Contopoulos-Ioannidis and Lau, 2001; Shadish, Clark and Steiner, 2008). The focus has been on the comparison of empirical results from observational studies and RCTs rather than theoretical considerations, as indicated in the literature (Ioannidis et al., 2001, for example). On the one hand, results in favour of observational studies were published (Concato, Shah and Horwitz, 2000): no systematic bias of results from observational studies could be substantiated compared to those from RCTs (Britton, McPherson, McKee, Sanderson, Black and Bain, 1998, for example), similar effect estimates were found, and the results were found to be less heterogeneous than those from RCTs (Concato et al., 2000).

On the other hand, Ioannidis et al. (2001) showed in an empirical evaluation that discrepancies beyond chance may occur and differences in estimated treatment effects are very common. In the effort to evaluate possible reasons for discrepant effect estimates between observational studies and RCTs other than randomisation, observational data were analysed by mimicking the design of a similar RCT as close as possible (Hernan, Alonso, Logan, Grodstein, Michels, Willett, Manson, and Robins, 2008) or the design of the study was chosen similar to the doubly randomised preference trial (Shadish et al., 2008).

The challenge arises in case where the model cannot account for all possible variations (Britton et al., 1998; McKee et al., 1999; Pocock and Elbourne, 2000). In another words as long as the uncertainty about missing or incompletely measured important covariates is existent or the representativeness of a study is not warranted, neither RCTs nor observational studies can properly clarify whether that what we estimate as the causal effect of intervention in the study is exactly that what we had in mind to estimate (Britton et al., 1998; Hernan et al., 2008).

2.2 Description of Causal Inference Problem

Among others, Schafer and Kang (2008) describe the problem of causal inference as follows: Consider a treatment that is either present or absent for each participant. The objective is to assess the average effect of the treatment on a subsequently measured outcome. We are likely to have a pretest (baseline) measure of the outcome and perhaps other variables that have been measured prior to treatment.

We assume for simplicity that there are no missing values in the baseline measures, that there is no dropout prior to final outcome, that the treatments have been carried out as intended with full compliance, and that there is no interference among participants in the sense that the treatment received by one has no effect on the outcome of any other. Scientists widely agree that to establish a causal link, it is not sufficient to show a significant difference in average response for treated and untreated persons at the end of the study. One must also rule out the possibility that the discrepancy is due to systematic differences between the groups at baseline.

If the treatment was randomly assigned as part of a designed experiment, that conclusion would be immediate, because randomisation will, on average, eliminate those differences. If the assignment was beyond the researchers' control, however, the groups may not have been equivalent at the outset, and ruling out alternative explanations is challenging and controversial. In another words, in the absence of randomisation, as in observational studies, one needs to be careful about the potential confounders in estimating the treatment effects (Lu, 2005).

Schafer and Kang (2008) reiterated the fact that popular strategy for ruling out alternatives is to measure as many confounders, pre-treatment variables that may be related to both the treatment and the response, as possible and then estimate what the difference in average response between treated and untreated persons would be if the average values of the confounders in both groups were equal. This idea, which underlies classical analysis of covariance (ANCOVA) and regression, still prevails in many areas of biomedical sciences, and social and behavioural sciences.

To appreciate how these methods work, it is helpful to understand the connections between causal inference and missing data. Notions of causality pertain to how an intervention would have changed individuals' results. With two different treatments, we can imagine two possible responses for each participant. Causal effects may be defined as differences between these so-called potential outcomes (Rubin, 1974; Letsoalo and Lesaoana, 2012). Because only one outcome can be observed for any participant, techniques for causal inference are, in essence, missing-data methods (Gourieroux and Monfort, 1981).

Here the focus is on the effect of a binary treatment on the mean of a numeric outcome. In practice, a treatment variable may be nominal, ordinal, or numeric, and the response could be any of these types as well. We also make the simplifying assumption that *all confounders have been measured and are available to the analyst* (Schafer and Kang, 2008). This assumption may approximately hold if an extensive set of measures known by subject-matter experts to be predictive of the treatment has been collected at the pretest (Schafer and Kang, 2008).

In other applications, this assumption will be highly questionable, especially if only a few demographic variables are present. Even under the assumption of unconfoundedness, causal inference is not trivial; many solutions have been proposed, and there is no consensus among statisticians about which methods are best (Schafer and Kang, 2008). We note that the best available answers under an assumption of no unmeasured confounders will establish useful benchmarks and points of departure for sensitivity analyses (Rosenbaum, 2005).

2.3 Rubin Causal Model: The Potential-Outcomes Framework for Causal Inference

Consider the classic case of a person who is treated at time t and, the outcome or response to the treatment is observed at time $t + k$ $(k > 0)$. How does one conclude that the treatment is effective or not? In another words, how do we measure the possible causal effect of the treatment? Donald Rubin's answer to estimating the causal effect of treatments in randomised and nonrandomised studies is based on a counterfactual proposition. A simple illustration would be *if an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone* (Rubin, 1974). Following Rubin's notation, if T represents taking two aspirins and C drinking just a glass of water, the potential outcomes Y relating to these two treatments may be written as two random variables, namely $Y(T)$ and $Y(C)$. The causal effect of the T versus C treatment on Y for a particular subject j observed or treated at time t and observed at time $t + k$ is then defined as $Y_j(T) - Y_j(C)$, i.e. the differential headache response to taking the aspirins or not taking them (Letsoalo and Lesaoana, 2012).

A researcher who wishes to estimate the effect of a treatment that she or he can control in an outcome of interest typically designs an experiment in which subjects are randomly assigned to alternative treatment and control groups. We note that other types of experiments are possible but randomised experiments are the most common research designs when researchers have control over the assignment of the treatments. After randomisation process, the experiment is run and the values y_j of observations in the treatment arms are recorded.

The mean difference in the observed outcomes across the study arms is called the estimated average effect (Hernan, 2004), and any ensuing debate then moves on to the particular features of the experimental protocol and the degree to which the pool of study participants reflect the target population for which one would wish to know the average treatment effect (Morgan and Whinship, 2007).

The definition of causal effect, (Morgan and Whinship, 2007; Letsoalo and Lesaoana, 2012) is given as:

Definition *Causal effect* of one treatment, say exposure T, over another, say control C, for a particular unit and an interval of initial time t_1 to final time t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to T initiated at t_1 and what would have happened at time t_2 if the unit had been exposed to C initiated at time t_1.

The assumptions are that

(a) a time of initiation of treatment can be ascertained for each unit exposed to T and C, and

(b) T and C are exclusive of each other in the sense that a treatment cannot simultaneously be a T and a C treatment.

If $Y(T)$ and $Y(C)$ are the values of Y measured at time t_2 on the units, given that the units received T and C, respectively, initiated at t_1 then the difference $Y(T) - Y(C)$ is called the causal effect (Letsoalo and Lesaoana, 2012).

In the actual world, one never observes both $Y(T)$ and $Y(C)$ at the same time for the same individual. The subject either takes (or is assigned to) T or C. Thus one can never observe for the same individual j the causal effect $Y_j(T) - Y_j(C)$. In general, people are assigned either to T or to C but not to both at the same time. Well, the outcome under treatments not assigned can be regarded as *missing*, and the problem is one of drawing inferences about these $missing$ values with the observed data. Basically, the absolute difference between treatments, measuring $Y(T) - Y(C)$, as well as the relative difference, measuring $Y(T)/Y(C)$, can be compared.

Given that it is impossible to calculate individual-level causal effects, then the attention is directed to the estimation of aggregated or average causal effects (see Tables 2.2 and 2.3). We need the stable unit treatment value assumption (SUTVA); and this assumption assumes that (Gelman and Hill, 2007, Page 178):

(a) the treatment status of any study unit j does not affect the potential outcomes of the other units (non-interference), and

(b) the treament for all units are comparable (no variation in treatments).

That is, we require that *the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units*. SUTVA implies that Y(T) and Y(C) (the potential outcomes for unit j) are in no way dependent on the treatment status of any other unit in the dataset. We immediately note that SUTVA is not just statistical independence between units!

Under SUTVA, the average treatment effect in the population is given by the expectation of the quantity δ_i as:

$$\begin{aligned} E(\delta) &= E[Y(T) - Y(C)] \\ &= \frac{1}{N} \sum_{1 \le j \le N} \Big[Y_j(T) - Y_j(C) \Big] \end{aligned} \tag{2.3.1}$$

where $E[\cdot]$ is the expectation from probability theory.

Also, we note that the expectation of the difference equals the difference of the two expectations (Hollard, 1986; Morgan and Whinship, 2007). Equation (2.3.1) reveals that information on different units (individuals) that can be observed can be used to gain knowledge about the average causal effect $E[\delta]$. The subscripting on j for δ_j has been dropped for (2.3.1). But this does not necessarily imply that σ is constant in the population, as is a random variable just like $Y(t), t = C$ or T. The subscript j can be dropped in (2.3.1) because the causal effect of a randomly selected individual from the population equals the average causal effect across individuals in the population. The average causal effect is the most common subject or quantity of investigation in biomedical sciences. Two conditional average treatment effects are of particular interest. The average treatment effects for those who specifically take the treatment and those who typically do not take the treatment.

Rubin's (1974) solution is often called the potential outcome (or response) model. The two potential outcomes being in this simple case $Y_j(T)$ and $Y_j(C)$ for each j. Note that the causal effect may differ from one individual to the other, thus a *typical* causal effect is obtained as above by taking the average (or any other summary measure) of the individual causal effects. The potential outcome approach to causal inference extends the conceptual apparatus of randomised experiments to the analysis of non-experimental data, with the goal of explicitly estimating causal effects of particular treatments of interest (Rubin, 1974).

In a setup especially clinical trial (as an example of a controlled experiment), where one wishes to measure the use-effectiveness of device versus some other intervention on some condition's outcome (e.g., pregnancy outcome). Usually, $\frac{N}{2}$ participants would be assigned to either intervention, and their effects of interventions are compared. The differential responses would then be $Y(T)$ and $Y(C)$ versus $Y(T)$ or $Y(C)$ in addition to $Y(T)$ versus $YC)$.

Following Rubin (1974), suppose there are only two subjects under study, denoted by 1 and 2. The typical causal effect (as defined above in the counterfactual situation) would then be

$$\frac{1}{2}\left[Y_1(T) - Y_1(C) + Y_2(T) - Y_2(C)\right] \tag{2.3.2}$$

In the actual world, one would observe in a single study either

$$Y_1(T) - Y_2(C) \tag{2.3.3}$$

or

$$Y_2(T) - Y_1(C) \tag{2.3.4}$$

depending on whether subject 1 or subject 2 is assigned to T, and vice versa (subject 2 or subject 1 to C).

If treatments are randomly assigned to subjects, we are equally likely to observe the difference as given by (2.3.3) or (2.3.4). The expected difference in the outcome Y is then the average of equations (2.3.3) and (2.3.4):

$$\frac{1}{2}\left[Y_1(T) - Y_2(C)\right] + \frac{1}{2}\left[Y_2(T) - Y_1(C)\right] \tag{2.3.5}$$

It is easily seen that under randomisation, Equation 2.3.5 is equal to Equation 2.3.3. In other words, Equation 2.3.5 is an unbiased estimate of Equation 2.3.3.

Suppose now that subjects 1 and 2 respond similarly to the treatments T and C. In that case

$$Y_1(T) - Y_2(C) = Y_2(T) - Y_1(C)$$

and furthermore

$$Y_1(T) - Y_2(C) = Y_1(T) - Y_1(C)$$

or

$$Y_2(T) - Y_1(C) = Y_2(T) - Y_2(C)$$

In the situation of perfectly matched subjects with respect to the effects of the treatments, the observed causal effect is equal to the counterfactual causal effect. Results under randomisation or perfect matching can be extended from two subjects to N subjects. Randomisation and matching are therefore two approaches to measuring the causal effect in experimental and nonexperimental studies, though randomisation cannot often be used in the social sciences and perfect matching is hardly possible in practice.

In many actual situations in nonexperimental research, the assignment of units to the case and control groups is based on self-selection. Thus the assignment procedure is often not $ignorable$, in the sense that the likelihood of treatment on the one hand and the outcome on the other hand are not independent.

For example, if the sickest opts for the new treatment and the healthier for the older or standard one, the outcome (e.g. recovery) in the treatment group will be due both to the new drug and to the characteristics of the patients at onset. In this case, one must control as best as possible for the assignment factors which have an impact on the outcome. In the above example, one would try to control for the state of health of both groups at the beginning of the trial.

Morgan and Whinship (2007) state that early work of counterfactual model was on experimental design by Neyman (1923), and the counterfactual model for causal analysis of observational data was formalised by Rubin since early 70s (see Rubin, 1974). The counterfactual model is often referred to as the *potential outcomes framework*.

There are several primitives, concepts that are basic and on which we must build. The fundamental notion underlying our view of causality is tied to an action or treatment or intervention applied to a unit. Here follows definition of treatment.

Definition A *treatment*[1] *or intervention*[2] is an action that can be applied or withheld from that unit. Here, a *unit* is a physical object, firm, or person, or collection of persons such as a classroom or a market, at a particular point in time.

Associated with each unit and each treatment there are two potential outcomes, the values of an outcome variable Y when the treatment is applied and when it is withheld. The objective is to learn about the causal effect of the application of the treatment relative to its being withheld on the outcome.

The core of the counterfactual model for observational data is: Suppose that each experimental unit or individual in a target population can be exposed to two alternative

[1] The application of an agent, surgery, psychotherapy, etc, to a study unit.

[2] Intervention comes from the Latin *intervenire*, and it is the act of inserting one thing between others, like a person trying to help. Often an intervention is intended to make things better.

states of a cause. By experimental unit we mean a particular unit of study (e.g., learner) at a particular time, since the effect of the treatment on a unit may depend on when treatment is applied. Only a finite number of experimental units need be considered since no treatment will be applied into the infinite future (Rubin, 1978). Each state is characterised by a distinct set of conditions, exposure to which potentially affects an outcome of interest. These alternative causal states are referred to as alternative treatment. Therefore, by treatment we mean a series of well-defined distinct actions that can be applied to a unit of study. The typical examples of treatments are medical or surgical interventions on patients with health condition or disease such as coronary artery disease (Rubin, 1978).

The key assumption of the counterfactual framework is that each study subject in the target population has a potential outcome under each treatment state, even though each individual can be observed in only one treatment state at any point in time (Morgan and Whinship, 2007). Thus, counterfactuals are possible outcomes in different hypothetical states of the world.

From the definition of causal effect we note that

(a) the definition does not depend on which outcome is actually measured, and

(b) the causal effect is the comparison of outcomes at the same moment of time, where the time of application of the treatment precedes that of outcome.

2.3.1 The Fundamental Problem of Causal Inference

The definition of “cause” is complex and challenging (Rubin, 1991), but for empirical research, the idea of causal effect of an agent or treatment seems more straightforward and practically useful. To define a causal effect in an individual study subject i, let us assume that we want to assess the effect of an index treatment or exposure level (active arm) as compared to another treatment or exposure level (control arm) on an outcome of interest (which can be discrete or continuous; qualitative or quantitative). In counterfactual inference it is assumed (Greenland and Brumback, 2002) that:

(a) at the fixed time point of assignment, the individual i could have been assigned to both treatment levels; and

(b) the outcome of interest exists under both treatment levels.

Causal comparisons entail contrasts between outcomes in possible states defined so that only the presence or absence of the treatment varies across the states. Since

Table 2.1: Hypothetical complete data: Illustration of what the complete data might look like, if it were possible to observe both potential outcomes on each unit. For each pair, the observed outcome is displayed in boldface.

	Pre-treatment	Treatment	Potential		Treatment
Unit, i	Inputs	Indicator	Outcomes		Effect
	X_{ij}	Z_i	Y(C)	Y(T)	$Y(T) - Y(C)$
1	x_{1j}	0	**69**	75	6
2	x_{2j}	1	**111**	108	-3
⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	$x_{(n-1)j}$	1	112	**111**	-1
n	x_{nj}	0	**92**	102	10

$Y(T)$ and $Y(C)$ exist in theory for each study subject, an individual-level causal effect can be defined as some contrast between $Y(T)$ and $Y(C)$, usually the simple difference $Y(T) - Y(C)$, as indicated by Table 2.1. A formal definition of treatment effect as given by, among others, Hofler (2005) and Gelman and Hill (2007) follows:

Definition The *treatment effect*[3] $\Delta = Y(T) - Y(C)$ indicates that unit i gains the increase/decrease in the response variable by Δ due to treatment. That is, the causal effect of one treatment relative to another for a particular experimental unit is the difference between the results; if instead, the unit had been exposed to a second treatment.

A major failing of the potential outcome model is that it cannot take *attributes* into account, e.g., gender is a cause of initial salary discrimination in many countries, ethnicity is a cause of differential HIV prevalence in Sub-Saharan Africa (Rubin, 1986). These attributes are not only associated with their respective effects - they are part of the causal mechanism itself. Any explanatory framework, especially in social sciences, that cannot take attributes into account is therefore necessarily flawed (Rubin, 1986).

Clearly, both observed and unobserved random variables are relevant. Causal effects cannot be observed or directly calculated at the individual-level since it is impossible to observe both $Y(T)$ and $Y(C)$ for any study subject (Hollard, 1986; Gelman and Hill, 2007; Morgan and Whinship, 2007). Accordingly, $Y(T)$ and $Y(C)$ are unobserved counterfactual outcome for each individual i in the control and treatment groups, respectively. Table 2.2 attempts to make this explanation more explicit, and Table

[3] Simply put - the causal effect of a treatment on an outcome for an observational or experimental unit i can be defined by comparing between the outcomes that would have occured under each of the different treatment possibilities.

Table 2.2: Illustration of the fundamental problem of causal inference. For each unit, we have observed some pre-treatment inputs, and then the treatment ($Z_i = 1$) or control ($Z_i = 0$) is applied. We can then observe one of the potential outcomes, (Y(C), Y(T)). As a result, we cannot observe the treatment effect, Y(T) - Y(C), for any of the units.

	Pre-treatment	Treatment	Potential		Treatment
Unit, i	Inputs	Indicator	Outcomes		Effect
	X_{ij}	Z_i	Y(C)	Y(T)	$Y(T) - Y(C)$
1	x_{1j}	0	y_{1c}	?	?
2	x_{2j}	1	?	y_{2t}	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	$x_{(n-1)j}$	1	$?$	$y_{(n-1)t}$	?
n	x_{nj}	0	y_{nc}	?	?

2.3 as given by Letsoalo and Lesaoana (2012), is a presentation of this information in a more compact format. Among others, Hollard (1986) describes this challenge as *the fundamental problem of causal inference* [4].

Table 2.3: The fundamental problem of causal inference

Intervention	Potential Outcome	Counterfactual Outcome	Causal Effect
Treatment	$Y(T)$	$Y(C)$	$Y(T) - Y(C)$
Control	$Y(C)$	$Y(T)$	

For a binary case with two causal states, labeled treatment and control, and associated potential outcome variables $Y(T)$ and $Y(C)$ we define a causal exposure variable, D, which takes on two values such that (Morgan and Whinship, 2007, Page 35):

$$D = \begin{cases} 1 & : \text{Treatment} \\ 0 & : \text{Control}. \end{cases}$$

Then, the observable outcome Y is given by:

$$Y = \begin{cases} Y(T) & : D = 1 \\ Y(C) & : D = 0. \end{cases}$$

This pair definition is written compactly as:

[4] That is, the so-called *fundamental problem of causal inference* is that atmost one of these potential outcomes, $Y(C)$ and $Y(T)$, can be observed for each study unit i.

$$Y = D{\cdot}Y(T) + (1 - D){\cdot}Y(C). \tag{2.3.6}$$

Equation (2.3.6) says one can never observe the potential outcome under the treatment state for those observed in the control state, and one can never observe the potential outcome under the control state for those observed in the treatment state (Morgan and Whinship, 2007).

Estimating causal effects requires one or some combination of the following (Green and Aronow, 2011; Gelman and Hill, 2007, Page 171):

(a) close substitutes for the potential outcomes,

(b) randomisation, and/or

(c) statistical adjustment.

In theory, the simplest solution to the fundamental problem of causal inference is to randomly sample a different set of units for each treatment group assignment from a common population, and then apply the appropriate treatment to each group. An equivalent approach is to randomly assign the treatment conditions among a selected set of units (Gelman and Hill, 2007).

Either of these approaches ensures that, on average, the different treatment groups are *balanced* or that the average observations in the different treatment conditions from the sample are estimating the average outcomes under control and treatment for the same population (Gelman and Hill, 2007).

In practice, however, we often work with observational data since observational studies can be more practical to conduct and can have more realism with regard to how the program or treatment is likely to be ‘administered’ in practice (Gelman and Hill, 2007, Page 181). However, results of clinical trials or experiments are subject to dispute, and in observational studies one basis for dispute is that since the study subjects were not randomly assigned to treatments, subjects at *greater risk* may be over-represented in some treatment groups.

We note that in observational studies treatments are observed rather than assigned, and it is reasonable to consider the observed data under treatment conditions as nonrandom samples from a common population. Therefore, in an observational study, there can be systematic differences between groups of study units that receive different treatments - differences that are outside the control of the practitioner who performs the experiment or trial[5] - and they can affect the outcome, say y. In this case

[5] Trial, experiment, test imply an attempt to find out something or to find out about something.

we need to rely on more data than just treatments and outcomes and the envisaged strategy for analysis has to rely upon stronger assumptions (see assumption (2.3.7) on Page 53).

2.3.2 The Assignment Mechanism

The four formal modes of causal inference that are considered conceptually distinct in Rubin (1990) are:

(1) Randomisation-based tests of sharp null hypotheses;

(2) Bayesian predictive inference for causal effects or, more descriptively, perhaps-full probability modelling for causal effects;

(3) Repeated-sampling randomisation-based inference; and

(4) Repeated-sampling model-based inference.

Detailed discussion of these modes is in Rubin (1990). A more important message than their differences, however, is that all modes share a common conceptual framework in which causal inferences can be drawn, and that a clear formulation of this framework is an essential ingredient of a valid statistical inference in a practical problem (Rubin, 1990). In particular, this common framework requires the specification of a posited assignment mechanism describing the process by which treatments were assigned to units; it is required for each mode of inference in the sense that causal answers generally change if the posited assignment mechanism is changed (Rubin, 1990, 1991). Rubin (1991, Page 1214) emphasised that the major source of uncertainty in the analysis of an observational study is generally not the mode of inference but rather the specification of this assignment mechanism.

Central to the potential outcomes framework is the assignment mechanism, the mechanism that determines which units get which treatment (Rubin, 2000, 2004).

Formally, the assignment mechanism is defined as a function assigning probabilities to all possible N-vectors of binary assignment Z given the N-vectors of potential outcomes $Y(C)$ and $Y(T)$ and the $N \times K$ matrix of covariates, $\mathbf{X}$, with i^{th} row equals $\mathbf{X}_i = (X_{1i}, \cdots, X_{1K})$, a K-vector of background variables which encodes characteristics of unit i (Rubin, 1991; Imbens and Rubin, 2000; Mattei, 2004). Mattei (2004, Page 7) provided the following definition:

Definition *Given a population of N units, the assignment mechanism is a row-exchangeable function $Pr\Big(\mathbf{Z}|\mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)\Big)$ taking on values in $\{0,1\}^N$ satisfying*

$$\sum_{\mathbf{Z}} Pr(\mathbf{Z}|\mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)) = 1,$$

for all $\mathbf{X}, \mathbf{Y}(C)$*, and* ***Y***$(T)$.

Randomised trial is an obvious example of an assignment mechanism (Rubin, 1991; Austin, 2011). It is an assignment mechanism such that

(i) it is ignorable, which means that it does not depend on the missing outcomes;

(ii) it is probabilistic, that is,

$$0 < Pr(Z_i = T|X, Y(C), Y(T)) < 1$$

for all i, and for all $\mathbf{X}, \mathbf{Y}(0)$, and $\mathbf{Y}(T)$, where

$$Pr(Z_i = T|\mathbf{X}, \mathbf{Y}(T) = \sum_{\mathbf{Z}:Z_i} Pr(\mathbf{Z}|\mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T))$$

is the unit assignment probability for unit i; and

(iii) it is a known function of its arguments. Here, we are mainly concerned with a special case of randomised experiments, classical randomised experiments, which in addition to the conditions required for randomised experiments assume local independence. This assumption requires the assignment mechanism to be separable in the unit assignment probabilities, at least conditional on $(\mathbf{Z}, \mathbf{X})$.

Moreover, it requires the unit assignment probability for unit i to be a function of outcomes and covariates for unit i only, free of the values of outcomes and covariates for other units other than through the dependence of the joint assignment probabilities on these outcomes (Mattei, 2004). Appendix C1 on Page 152 explains assignment as locally independent [see (Rubin, 1991, Page 1220)].

The assignment mechanism is fundamental to causal inference because it tells us how we got to see what we saw. Causal inference is basically a missing data problem[6] because at least half of the potential outcomes are not observed, and hence missing (Rubin, 1977, 2007; Mattei, 2004). Without understanding the process that creates missing data, we have no hope of inferring anything about them (Hernan, 2004). Without a model for how treatments are assigned to individuals, formal causal inference, at least using probabilistic statements, is impossible (Rubin, 1991). This does not mean that we need to know the assignment mechanism, but rather that without positing one, we cannot make any statistical claims about causal effects, such as unbiasedness of estimates, confidence coverage of intervals for effects, significance levels of tests, or coverage of Bayesian posterior interval (Mattei, 2004; Morgan and Rubin, 2012).

[6] Carter (2006) discusses solutions for missing data in structural equation modeling (SEM).

2.3.3 Causal Effects in Randomised Experiments

A randomised controlled trial (RCT) randomises who receives a program (or service) – the study Arm1 (say treatment group) - and who does not – the study Arm 2 (say, the control). It then compares outcomes between those two groups; this comparison gives us the impact of the program. Figure 2.1 attempts to explain this process .

Randomised experiments provide the analyst with the opportunity to achieve unbiased estimation of causal effects. Figure 2.1 is a schematic representation of two-arms study. Unbiasedness is an important statistical property, entailing that the expected value of an estimator is equal to the true parameter of interest. Randomised experiments are often justified by the fact that they facilitate unbiased estimation of the average treatment effect (ATE) (Aronow and Middleton, 2013).

However, one of the major constraints to any quantitative impact evaluation – not just RCTs – is sample size. In the case of RCTs, a concern is with sample size along two dimensions: the unit of analysis, and the unit of randomisation. Both the unit of analysis and the unit of randomisation are integral in determining statistical significance and statistical power.

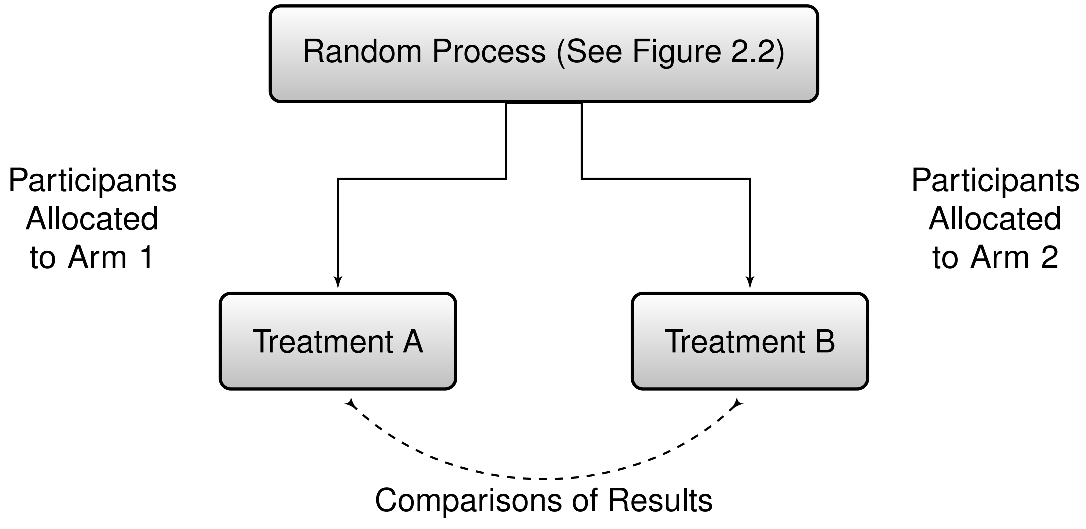

Random Process (See Figure 2.2)
Participants Allocated to Arm 1
Participants Allocated to Arm 2
Treatment A
Treatment B
Comparisons of Results

Figure 2.1: Randomisation Trial
Source: Letsoalo (2004, Page 15)

This unbiasedness is undermined when the analyst uses an inappropriate analytical tool (Suresh, 2011; Aronow and Middleton, 2013). Randomisation is usually associated with trial blindness (Saghaei, 2011; Aronow and Middleton, 2013).

Randomisation process eliminates the selection bias, balances the study groups with respect to many known and unknown confounding or prognostic variables, and forms the basis for statistical tests, a basis for an assumption of free statistical test of the equality of treatments. In general, a randomised experiment is an essential tool for testing the efficacy of the intervention (Saghaei, 2011; Suresh, 2011).

The gold standard for the estimation of causal effects is to conduct randomised experiments, such as clinical trials (Trojano, Pellegrini, Paolicelli, Fuiani and Di Renzo, 2009; Morgan and Rubin, 2012; Fives, Russell, Kearns, Lyons, Eaton, Canavan and O'Brien, 2013). Simply put - Randomised controlled trials (RCTs), or randomised impact evaluations (RIE), are a type of impact evaluation which uses randomised access to social programs as a means of limiting bias and generating an internally valid impact estimate.

In another words, the randomised controlled trials are widely accepted as the most powerful research methods for minimising bias when evaluating interventions e.g., health technologies (Morgan and Rubin, 2012; Fives et al., 2013). Randomisation is an assignment mechanism that allows particulary straightforward estimation of causal effects (Suresh, 2011; Suresh and Chandrashekara, 2012; Aronow and Middleton, 2013).

> *Randomisation is the process of making something random; in various contexts. This involves, for example, generating a random permutation of a sequence (such as when shuffling cards, flipping a coin or rolling a pair of [balanced] dice). Randomisation is usually associated with trial blindness (Saghaei, 2011; Aronow and Middleton, 2013).*

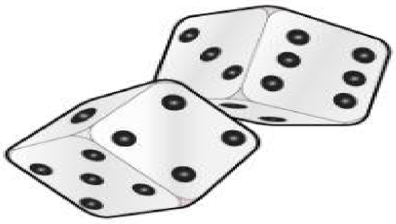


Figure 2.2: Rolling of a pair of balanced dice: Generating an assignment process

Therefore, simple randomised experiments form the basis for inference for causal effects in more complicated situations, such as when the assignment probabilities depend on covariates or when there is noncompliance with the assignment mechanism (Mattei, 2009). In addition, an unconfounded assignment mechanism, which essentially is a set of randomised experiments, forms the basis for the analysis of an observational nonrandomised study by using the randomised experiment as a template (Mattei, 2009).

Proper randomisation ensures no a priori knowledge of group assignment (i.e., allocation concealment). That is, researchers, subject or patients or participants, and others should not know to which group the subject will be assigned. Knowledge of group assignment creates a layer of potential selection bias that may taint the data (Suresh, 2011). An attempt to answer the question concerning the benefits of randomisation in determining the causal effect of the active treatment versus control treatment on an outcome Y is entertained. Therefore, suppose that a randomised experiment with N trials has been performed to estimate the typical causal effect of the active versus control treatment on Y for some population of units (assuming no pre-treatment covariates have been recorded).

Mattei (2004) argues that randomisation can never assure us that we are correctly estimating the causal effect of one treatment versus another for the N trials under study, but it provides important benefits besides the intuitive ones that follow from making all systematic source of bias into random ones. Randomisation provides a mechanism to derive probabilistic properties of estimates without making other assumptions (Rubin, 1974). For example, a special case of randomisation is application in randomised controlled trials. Randomised controlled trials are the most rigorous way of determining whether a cause-effect relation exists between treatment and outcome and for assessing the cost effectiveness of a treatment. RCTs have several important features:

i. Random allocation to study groups,

ii. Study participants and trialists should remain unaware of which treatment was given until the study is completed-although such double blind studies are not always feasible or appropriate,

iii. All intervention groups are treated identically except for the experimental treatment,

iv. Study participants are normally analysed within the group to which they were allocated, irrespective of whether they experienced the intended intervention (intention to treat analysis), and

v. The analysis is focused on estimating the size of the difference in predefined outcomes between interventionb or study groups.

The two important properties of randomisation are (Mattei, 2004; Aronow and Middleton, 2013):

(1) the average difference between the treatment and control group is an unbiased estimate of τ, the typical causal effect for the N units in the study, as given or defined by Equation (2.3.1); and

(2) precise probabilistic statements can be made indicating how unusual the observed average difference between the treatment and control group would be under specific hypothesised causal effects.

Given that all study participants are randomly allocated to their respective study arms does not guarantee representativity or representativeness of the study participants. This implies that there is a challenge with regard to statistical inference[7]. Rubin (1974) urges analysts to make subjective random sampling assumption - in order to believe their findings are useful. The justification for the two study arms to be comparable is given by Appendix C2.

2.3.4 Causal Effects in Observational Studies

A major issue is that observational studies are more exposed and prone to biases than RCTs (Trojano et al., 2009). Many studies in social science that aim to estimate the effect of an intervention suffer from treatment selection bias (Klein-Geltink, Rochon and Dyer, 2007), where the units who receive the treatment may have different characteristics from those in the control condition (Stuart and Rubin, 2008). The fundamental criticism of observational studies, attempting to estimate the effect of a treatment by comparing outcomes for nonrandomised subjects, is that either known or unknown confounding factors[8] may influence the measured association between an exposure of interest and a given outcome. Table 2.4 presents a summarised source of bias as given by Trojano et al. (2009).

Differences in outcomes can be due to differences between the study groups, in ascertainment of outcomes, unintended differences in other treatment factors, or to the treatment factor being studied and even the outcome measures (Klein-Geltink et al., 2007; Trojano et al., 2009; Green and Aronow, 2011).

Bias[9] can also be a problem in observational studies when there is a difference in the reliability of the data collected on treatment exposure between cases that have the

[7] Trying to reach conclusions that extend beyond the immediate data alone - Inferential statistics

[8] Confounding variable (also confounding factor, a confound, or confounder) is an extraneous variable in a statistical model that correlates (directly or inversely) with both the dependent variable and the independent variable.

[9] *Bias* is an error in design or execution of a study, which produces results that are consistently distorted in one direction because of nonrandom factors. Bias can occur in randomised controlled trials but tends to be a much greater problem in observational studies. *Selection bias* is a distortion in the estimate of association between risk factor and disease (or condition) that results from how the subjects are selected for the study. Selection bias could occur because the sampling frame is

Table 2.4: Bias in observational studies of treatment

Source	Explanation
Confounding	Systematic error due to the failure to account for the effect of one or more variables that are related to both the causal factor being studied and the outcome and are not distributed the same between the groups being studied. Confounding occurs when a factor is associated with the use (confounding by indication) or avoidance (confounding by contraindication) of the treatment, but independently influences the risk of the outcome of interest.
Recall bias	Systematic error that occurs when the reliability of recall of treatment exposure differs between those who develop an adverse outcome and those who do not.
Detection bias	Systematic error that occurs when, because of the lack of blinding or related reasons, the measurement methods are consistently different between groups in the study.

charecteristic of interest and controls that do not. Among others, Klein-Geltink et al. (2007) underlined that important issues related to confounding are often not clearly addressed, from the reader's perspective, in published observational studies. Failure to recognise the limitations of observational studies in the assessment of treatment effects may have serious consequences, including both the use of ineffective or dangerous treatments and the inappropriate abandonment, or insufficiently widespread use of effective treatments (Klein-Geltink et al., 2007; Trojano et al., 2009).

An important problem of causal inference is how to estimate treatment effects in observational studies. It is well recognised that the estimate of causal effect obtained by comparing a treatment group with a nonexperimetal comparison group could be biased because of problems such as self-selection or some systematic judgements by the researcher in selecting units to be assigned to the treatment (Dehejia and Wahba, 2002). That is, the effect of variables not explicitly controlled is usually more serious in nonrandomised than in randomised studies; secondly, the applicability or generalisation of the results to a population of interest is often more serious in nonrandomised than in randomised studies. In the statistical analysis of observational data, propensity score matching (PSM) is a methodology attempting to provide

sufficiently different from the target population or because the sampling procedure cannot be expected to deliver a sample that is a mirror image of the sampling frame. *Information bias* is a distortion in the estimate of association between risk factor and disease that is due to systematic measurement error or misclassification of subjects on one or more variables, either risk factor or disease status. It is important to realise that these errors are part of being human and they are not occurring because the physicians or researchers are not being sufficiently careful. It is not so much the random mismeasure or misdiagnosis of an individual that is problematic (although random errors in diagnosis will tend to bias the association toward a relative risk of 1.0, because the true association is diluted with noise). It is the method of measurement or classification that is the greater problem, because it systematically exerts an effect on each of the individual measurements in the sample.

unbiased estimation of treatment-effects. The possibility of *bias* arises here because the effectiveness of a treatment may depend on characteristics that are associated with whether or not a participant in an observational study chooses, or is chosen, to receive a given treatment.

The point about using randomisation is that it avoids any possibility of selection bias in a trial. The test that randomisation has been successful is that different treatment groups have same characteristics at baseline. In another words, in randomised experiments, the results in the two treatment groups may often be directly compared because their units are likely to be similar, whereas in nonrandomised experiments, such direct comparisons may be misleading because the units exposed to one treatment generally differ systematically from the units exposed to the other treatment (Rosenbaum and Rubin, 1983b; D'Agostino, 1998). Specifically, whereas in experimental situations one can obtain a control and treatment group which are homogeneous with respect to the observable characteristics, $\mathbf{X}$, this is not possible in nonexperimental studies since it is likely that the decision to be assigned to a treatment is in this case not independent from the observable as well as unobservable characteristics (Mattei, 2009). Note that with random assignment, homogeneity of the control and treatment group with respect to the unobservable characteristics is also guaranteed if the size of the groups is sufficiently large (Donner and Klar, 2000).

2.3.5 Motivation for Matching

Matching is a common technique used to select control study participants who are *matched* with the treated study participants on background covariates that the researcher believes need to be controlled. Although the idea of finding matches seems straightforward, it is often difficult to find study participants who ae similar (that is, can be matched) on all important covariates, even when there are only a few background covariates of interest (D'Agostino, 1998).

Matching methods are commonly used in two types of settings which are the situation in which the outcome values are not yet available and matching is used to select subjects for follow-up and the second setting is the one in which all the outcome data is already available (Stuart, 2010). In this case the goal of the matching is to reduce bias in the estimation of the intervention effect (Kuehl, 2000; Stuart, 2010).

The effectiveness of the intervention is evaluated by the use of acceptable methods for comparing study groups. The usual research paradigm consists of the following method (Kuehl, 2000; Seeger and Walker, 2007):

(a) Form treatment and experimental groups, sometimes with a single group serving as its own control, such as in pre intervention and post intervention studies or crossover studies.

(b) Map treatments to groups.

(c) Analyse group differences.

(d) Generalise the findings based on groups to tendencies among future individuals.

Defining the groups is a critical first step. It is important to properly form groups, as research question/s might not be addressed adequately if groups are not comparable. Once the groups are defined, one would want the composition of the groups to be identical. Short of that ideal, statistical adjustments, often in the form of blocking variables or covariate analysis, could be used to adjust for the pre-treatment group differences (Kuehl, 2000; Seeger and Walker, 2007).

In a randomised drug trial, the allocation of subjects to treatment and control groups at random leads to groups that are similar with respect to both measured and unmeasured baseline characteristics[10]. This baseline comparability supports the conclusion that any differences between groups in the occurrence of outcomes during follow-up must be due to the one characteristic that differs between the two groups by design or allocation strategy (Seeger and Walker, 2007). As highlighted by Stuart (2010, Page 1), one of the key benefits of randomised experiments for estimating causal effects is that the study groups are guaranteed to be only randomly different from one another on all background covariates, both observed and unobserved.

In an observational setting, matching groups of patients or participants (cohorts) on the characteristics that are part of the prescribing decision creates a balance with respect to measured characteristics that are even more complete than what results from randomisation, so that treatment and control groups are identical at the start of follow-up (Kuehl, 2000).

By forming treated and untreated groups that are individually matched in this way, an observational study of intervention effect, e.g. educational intervention, is possible given two considerations:

(a) that the list of matching characteristics is complete (it includes all variables that actually went into the prescribing decision so that no unmeasured predictor of treatment is present), and

[10] A formal expression for consequences of random allocation is that treatment allocation, being based on a random number, is uncorrelated with any possible study participant characteristic. At least in expectation and in fact with large numbers, treatment allocation being uncorrelated with participant characteristics means that the distribution of characteristics is the same in all treatment groups (Seeger and Walker, 2007).

(b) that there is a control individual with each collection of attributes that can be matched to each individual who received the drug.

In a real-world setting, this form of exhaustive matching is difficult because these considerations are in conflict. Many characteristics can plausibly enter into the prescribing decision, leading to such a large number of combinations of characteristics that it becomes impossible to find a control individual with exactly the same characteristics as each treated individual. This problem is called *curse of dimensionality* (Seeger and Walker, 2007). Therefore, conditioning on all relevant covariates is limited in case of a high dimensional vector ($\mathbf{X}$, say).

There are two main approaches to allocation of study units to interventions in research or clinical trials. These are unrestricted allocations and restricted allocations. Under unrestricted allocation we have (completely) randomised allocation, in which study units from a single pool, with no pre-stratification or matching according to baseline characteristics, are allocated to treatment groups. We define treatments as set of circumstances created for the experiment in response to the research hypothesis, and they are the focus of research (Kuehl, 2000).

Under restricted[11] allocation, study units are first divided into strata according to baseline characteristics and then allocated to groups within the designated strata. The aim of restricted allocation, stratification or matching, is to provide groups that are more homogeneous or evenly balanced with respect to baseline characteristics (Kuehl, 2000). Statistical power will also be increased provided the baseline characteristics selected as stratifying factors are strong predictors of the outcome of interest (Pocock, Assmann, Enos and Kasten, 2002).

Matching techniques have origins in experimental work from the first half of the twentieth century. In the 1980s, matching techniques were advanced in a set of papers that offered solutions to a variety of practical problems that had limited matching technique to very simple applications in the past (Kuehl, 2000; Morgan and Whinship, 2007).

An alternative to random assignment is a matched-pairs design[12]. Each member of the first group is matched with a member of the second group on all the factors the researcher considers to be feasible and relevant. In a well-matched pair, it is as if we

[11] Any procedure used with random assignment to achieve balance between study groups in size or baseline characteristics (Kuehl, 2000). Blocking is used to ensure that comparison groups will be of approximately the same size.

[12] A matched pairs design is a special case of a randomised block design. It can be used when the experiment has only two treatment conditions; and subjects can be grouped into pairs, based on some blocking variable. Then, within each pair, subjects are randomly assigned to different treatments.

are using the same individual twice. When matching is adequate, the variables used for matching that might cause confounding problems are controlled (Pocock et al., 2002).

The approach falls apart when one matches on too few or irrelevant covariates (matching variables), as the match is not necessarily a good one. Matching on many covariates is difficult, especially if one is trying to obtain an exact match when some of the covariates are continuous.

In the methodological literature, matching is usually introduced in one of the two ways (Morgan and Whinship, 2007, Page 88):

(a) as a method to form quasi-experimental contrasts by sampling comparable treatments and control cases from among two larger pools of such cases, or

(b) as a nonparametric method of adjustment for treatment assignment patterns when it is feared that ostensibly simple parametric regression estimators cannot be trusted.

A possible way to address this complication in nonexperimental studies or in observational studies, is to consider the randomised experiment as a template for the analysis of an observational or nonrandomised study (Murry, Varnell and Blitstein, 2004). Having the template of a randomised experiment means having to think about the underlying randomised experiment that could have been done (Murry et al., 2004), where in the randomised experiment underlying an observational study, the probabilities of assignment to treatments are not equal, but are rather functions of the covariates, and so the template is actually an unconfounded assignment mechanism (Mattei, 2009).

Imbens (2004) recommended that the strong ignorability or unconfoundedness assumption is (highly) needed to accomplish this task. According to Imbens (2004) unconfoundedness assumption was first presented in this form, as given by Expression (2.3.7), by Rosenbaum and Rubin (1983b), who refered to it as “ignorable treatment assignment”. Generally, it is said treatment assignment is strongly ignorable given a vector of covariates $\mathbf{X}$ if (Rosenbaum and Rubin, 1983b):

$$(Y(C), Y(T)) \perp Z | \mathbf{X} \tag{2.3.7}$$

and

$$0 < Pr(Z = T|\mathbf{X}) < 1, \tag{2.3.8}$$

for all $\mathbf{X}$.

The strong ignorability assumption asserts that the probability of assignment to a treatment does not depend on the potential outcomes conditional on observed covariates. Simply put: the distribution of the potential outcomes is the same across levels of the treatment Z once we condition on confounding covariates $\mathbf{X}$ (Mattei, 2004). In another words, within subpopulations defined by values of the covariates, we have random assignment. The ignorability of treatment assignment is the so-called in statistics literature and *selection on observables* in econometrics (Mattei, 2004).

The ignorability assumption rules out the role of the unobservable variables (Manski, Sandefur, McLanahan and Powers, 1992). The issue of unobserved covariates should be addressed using models for sensitivity analysis (Rosenbaum and Rubin, 1983a) or using nonparametric bounds for treatment effects (Manski, 1990; Manski et al., 1992).

Clearly, the strong ignorability assumption may be controversial. It requires that all variables that affect both outcomes and the likelihood of receiving the treatment are observed. Although this is not testable, it clearly is a very strong assumption, and one that need not generally be applicable. We view it as a useful starting point for two reasons (Mattei, 2009).

First, even if these attempts are not completely successful the assumption that all relevant variables are observable may be a reasonable approximation, especially if much information about pre-treatment outcomes is available. Second, any alternative assumption that does not rely on unconfoundedness while allowing for consistent estimation of the average treatment effects must make alternative untestable assumptions, such as the instrumental variable technique (Angrist, 1990; Angrist and Krueger, 1991). Whereas the unconfoundedness assumption[13] implies that the best instrument variable technique matches units that differ only in their treatment status, but otherwise are identical. Alternative assumptions implicitly match units that differ in the pre-treatment characteristics.

The unconfoundedness assumption therefore may be a natural starting point after

[13] Unconfoundedness assumption implies that adjusting for differences in observed pretreatment variables removes biases from comparisons between treated and control units. This assumption is not directly testable (Imbens, 2004).

comparing average outcomes for treated and control units to adjust for observable pre-treatment differences (Hirano and Imbens, 2001). The unconfoundedness assumption validates the comparison of treated and control units with the same value of covariates. The treatment effect for the subpopulation with $\mathbf{X} = \mathbf{x}$ can be written as (Imbens, 2004):

$$\begin{aligned}\sigma(X) &= E(Y(T) - Y(C)|\mathbf{X} = \mathbf{x}) \\ &= E(Y(T)|Z = T, \mathbf{X} = \mathbf{x}) - E(Y(C)|Z = C, \mathbf{X} = \mathbf{x}) \\ &= E(Y|Z = T, \mathbf{X} = \mathbf{x}) - E(Y|Z = C, \mathbf{X} = \mathbf{x})\end{aligned} \tag{2.3.9}$$

where both terms on the right-hand side of Equation (2.3.9) can be estimated from a random sample of ($\mathbf{X}$,Z,Y). The average treatment effect can then be estimated using the equality

$$\sigma = E(\sigma\mathbf{X}). \tag{2.3.10}$$

2.4 Propensity Score

Several possible methodological improvements (regression adjustment, stratification and matching) have been proposed and are available to deal with confounding (Murry,Varnell and Blitstein, 2004) and to improve validity when randomisation is absent (Murry et al., 2004; Trojano et al., 2009; Austin, 2011).

(a) Regression analyses estimate the association of each independent variable (baseline characteristics and the intervention) with the dependent variable (outcome of interest) after adjusting for the effects of all the other variables, so that they provide an adjusted estimate of the intervention effect (Trojano et al., 2009).

(b) Stratification consists of grouping subjects into strata determined by observed background characteristics believed to confound the analysis. Treated and control subjects in the same strata are compared directly. Stratification creates subgroups that are more balanced in terms of confounders than the total population which can result in less biased estimates of the intervention effect. The comparisons from different strata are then combined to give a final estimate of intervention effect (Gelman and Hill, 2007).

(c) Matching techniques allow to match individual cases (i.e. treated patients) with individual controls that have similar confounding factors in order to reduce the effect of these on the association being investigated in analytical studies. This is most commonly seen in case-control studies and when there are only limited numbers of treated patients and a much larger number of untreated (or control) patients (Kuehl, 2000; Trojano et al., 2009).

However, these traditional methods of adjustment are often limited since they can only use a small number of covariates for adjustment or if there is extreme imbalance in the background characteristics (D'Agostino, 2007).

Typically, there are many background characteristics that need to be controlled for estimating the average causal effect σ, and adjusting the estimation for all these covariates can be actually infeasible (*curse of dimensionality*) (Rubin, 1997). To overcome this challenge, the use of the balancing scores $b(X)$, i.e. functions of the relevant observed covariates X such that the conditional distribution of X given $b(X)$ is independent of assignment into treatment was suggested (Rubin, 1997; Rosenbaum and Rubin, 1983b).

One possible balancing score is propensity score technology, introduced by Rosenbaum and Rubin (1983b). Propensity score reduces the entire collection of background characteristics to a single composite characteristic that appropriately summarises the collection (Rosenbaum and Rubin, 1983b). It is an alternative way of dealing with confounding caused by nonrandomised assignment of treatments in cohort studies. The following is the definition of propensity score, as given by among others, Rosenbaum and Rubin (1983a,b).

Definition *Propensity Score* $e(\mathbf{X})$ is the conditional probability of receiving a treatment given pre-treatment characteristics such that $e(\mathbf{X}) = Pr(Z = T|\mathbf{X})$.

All pre-intervention or pre-treatment covariates need to be controlled; for they are all confounding covariates. Table 2.5 attempts to make the definition of propensity score more explicit. The propensity score is a balancing score, that is treatment assignment and observed covariates are conditionally independent given the propensity score:

$$\mathbf{X} \perp Z | e(\mathbf{X}). \tag{2.4.1}$$

In particular, the propensity score is the coarsest balancing score, i.e., any balancing score b(X) must satisfy the relation $e(\mathbf{X}) = f(b(\mathbf{X}))$, for some function $f(\cdot)$ (Rosenbaum and Rubin, 1983a).

Among others, Rosenbaum and Rubin (1983a,b); Ichimura and Taber (2001); Imbens (2004); Mattei (2004); Dehejia (2005); Abadie and Imbens (2006); Morgan and Whinship (2007); Austin (2009) and Mattei (2009) have stressed the fact that the

Table 2.5: Propensity score

Definition	Device for balancing numerous observed covariates
Formal	The conditional probability of exposure to a treatment given observed covariates.
Intuitive	The likelihood that a person would have been treated using only their covariate scores.
Comment	Collection of covariates is collapsed into a single variable; the probability (or propensity) of being treated.
Limitation	Does not control for unobserved variables that may affect subjects that have received treatment.

key feature of propensity score methodology is that, given the strong ignorability assumption, treatment assignment and the potential outcomes are independent:

$$(Y(C), Y(T)) \perp Z | e(\mathbf{X}), \tag{2.4.2}$$

and

$$0 < Pr(Z = T | e(\mathbf{X})) < 1. \tag{2.4.3}$$

The advantage of adjusting for propensity score lies in the fact that it removes the bias associated with differences in the observed covariates in the study arms (D'Agostino, 1998). As a result, given the strong ignorability assumption, if the propensity score $e(X)$ is known, then Imbens (2004, Page 9) provided a proof that:

$$\begin{aligned}\sigma &= E(Y(T) - Y(C)) \\ &= E(E(Y(T) - Y(C)|e(\mathbf{X}))) \\ &= E\Big(E(Y(T)|Z = T, e(\mathbf{X})) - E(Y(C)|Z = C, e(\mathbf{X})))\end{aligned} \tag{2.4.4}$$

where the outer expectation is over the distribution of $e(\mathbf{X})$.

The propensity score is a potential matching variable because it does not depend on response information that will be collected after matching (Rosenbaum and Rubin, 1985). Since exact matching for a known propensity score will on average remove all the bias in $\mathbf{X}$ (D'Agostino, 1998), the propensity score $e(\mathbf{X})$ is in a sense the most important scalar matching variable (Rosenbaum and Rubin, 1985).

Matching on $e(\mathbf{X})$ balances the observed covariates $\mathbf{X}$; however, unlike randomisation, matching on $e(\mathbf{X})$ does not balance unobserved covariates except to the extent that they are correlated with $\mathbf{X}$: we need the strong ignorability assumption[14].

In practice, several issues need to be addressed before the propensity score can be used as a matching variable. First, the functional form of $e(\mathbf{X})$ is rarely if ever known, and therefore $e(\mathbf{X})$ must be estimated from the available data. Second, exact matches will rarely be available, and so issues of closeness on $e(\mathbf{X})$ must be addressed. Third, adjustment for $e(\mathbf{X})$ balances $\mathbf{X}$ only in expectation, that is, averaging over repeated studies (Rosenbaum and Rubin, 1985). In any particular study, further adjustments for $\mathbf{X}$ may be required to control chance imbalances in $\mathbf{X}$. Such adjustments, for example, by covariance analysis, are often used in randomised experiment to control chance imbalances in observed covariates.

2.4.1 How Matching Methods are Implemented

The goal of matching[15] is to create a dataset that looks closer to one that would result from a perfectly blocked (and possibly randomised) experiment (Kuehl, 2000; Stuart, 2010). A crucial part of any matching procedure is, therefore, to assess how close the (empirical) covariate distributions are in the two groups, which is known as *balance* (Ho, Imai, King and Stuart, 2011). Because the outcome variable is not used in the matching procedure, any number of matching methods can be tried and evaluated, and the one matching procedure that leads to the best balance can be chosen (Ho et al., 2011).

Matching methods have four key steps, with the first three representing the *design* and the fourth the *analysis* (Stuart, 2010, Page 5). The four key steps used for matching methods when estimating causal effect are (Stuart and Rubin, 2008; Stuart, 2010):

1. Defining “closeness”: the distance measure used to determine whether an individual is a good match for another,

2. Implementing a matching method, given that measure of closeness,

3. Assessing the quality of the resulting matched samples, and perhaps iterating with Steps (1) and (2) untill well-matched samples result, and

[14] Rosenbaum and Rubin (1983a) and Rosenbaum (1984) discuss methods for addressing the possible effects of unobserved covariates in observational studies.

[15] Stuart and Rubin (2008) refer to matching as a quasi-experimental design. Matching is a special form of stratification in which there are constraints on the number of observed treated and control units in each stratum.

4. Analysis of outcome and estimation of the treatment effect, given the matching-done in Step 3.

Appendix D on Page 156 provides further information with regard to each of the mentioned steps.

2.4.2 How to Estimate Propensity Score

This section gives a brief description of how to estimate and use propensity score methodology in practical applications.

In general, exact matches on propensity score is impossible to obtain, so methods which seek approximate matches must be used (Mattei, 2004). Here follow some properties of some matching methods based on the propensity score.

Mattei (2004) gives the mean bias or expected difference in $\mathbf{X}$ prior and after to matching, respectively as

$$E(\mathbf{X}|Z = T) - E(\mathbf{X}|Z = C), \tag{2.4.5}$$

and

$$E(\mathbf{X}|Z = T) - E_M(\mathbf{X}|Z = C), \tag{2.4.6}$$

where $E_M(\mathbf{X}|Z = C)$ is the expected value of $\mathbf{X}$ in the matched control group. Generally, $E_M(\mathbf{X}|Z = C)$ depends on the matching method used, whereas $E(\mathbf{X}|Z = T)$ and $E(\mathbf{X}|Z = C)$ depend only on population characteristics (Mattei, 2004). A matching method is equal-percent bias reducing (EPBR) if the reduction in bias is the same for each coordinate of $\mathbf{X}$ (Rubin, 1976; Rosenbaum and Rubin, 1983b), that is, if

$$E(\mathbf{X}|Z = T) - E_M(\mathbf{X}|Z = C) = \gamma(E(\mathbf{X}|Z = T) - E(\mathbf{X}|Z = C)) \tag{2.4.7}$$

for some scalar $0 \leq \gamma \leq 1$ (Rubin, 1976). If a matching method is not EPBR, then matching actually increases the bias for some linear functions of $\mathbf{X}$. If little is known about the relationship between $\mathbf{X}$ and the response variables that will be collected after matching, then EPBR matching methods are attractive, since they are the only methods that reduce bias in all variables having linear regression on $\mathbf{X}$. Rosenbaum and Rubin (1983b) showed that matching on the population propensity score alone is EPBR whenever $\mathbf{X}$ has a linear regression on some scalar function of e; that is, whenever $E(X|e) = \alpha + \gamma' g(e)$ for some scalar function $g(\cdot)$.

Matched samples can be constructed by using several different methods that matched treated units to control units. Two standard techniques are the nearest available matching on the propensity score and "subclassification on the propensity score" (Austin, 2011). Subclassification method consists of dividing experimental and control units on basis of $e(\mathbf{X})$ into subclasses or strata such that within each subclass treated and control units have on average the same propensity score. Then, within each stratum in which both treated and control units are present, the average outcomes of the treated and control units are compared (Rosenbaum and Rubin, 1983b).

The average treatment effect of interest is finally obtained as an average of the subclass-specific comparisons. One of the pitfalls of the subclassification method is that it discards observations in strata where either treated or control units are absent (Mattei, 2004). This observation suggests an alternative way to match treated and control units, which consists of taking each treated unit and searching for the control unit with the closest propensity score, i.e., the nearest available matching on the propensity score (Mattei, 2004). Although it is not necessary, the method is usually applied with replacement, in the sense that a control unit can be a best match for more than one treated unit. Once each treated unit is matched with a control unit, the difference between the outcome of the treated units and the outcome of the matched units is computed. The average treatment effect of interest is then obtained by averaging these differences (Mattei, 2004).

Usually we do not actually know the propensity scores, and so we must estimate them. Propensity scores can be estimated in a number of different ways, including discriminant or CART analysis. Propensity scores can be estimated using standard models such as logistic regression, where the outcome is the treatment indicator and the predictors are all the confounding covariates (Gelman and Hill, 2007, Page 207). Then matches are found by choosing for each treatment observation the control observation with the closest propensity score (Gelman and Hill, 2007, Page 207). In principle, any standard probability model can be used to estimate the propensity score. For instance,

$$Pr(Z = T|\mathbf{X}) = F(h(\mathbf{X})), \tag{2.4.8}$$

where $F(\cdot)$ is the normal or the logistic cumulative distribution function and $h(\mathbf{X})$ is a function of covariates with linear or higher order terms.

The outcome variable plays no role in the estimation of the propensity score (Greenland, 2004), as such estimating propensity scores only involves the covariates. Greenland (2004) stress that the success of the propensity score estimation must be assessed by the resultant balance of the observed distribution of covariates across study groups. Therefore, the fit of the models used to create estimated propensity scores is discouraged if it is to be used for evaluation for success of the process of propensity score estimation (Greenland, 2004).

The goal of propensity score matching is not to ensure that each *pair of matched observations* is similar in terms of all their covariate values, but rather that the matched groups are similar on *average* across all their covariate values. Thus, the adequacy of the model used to estimate propensity score can be evaluated by examining the balance that results on average across the matched groups (Gelman and Hill, 2007, Page 207).

Models for the data, $Pr(\mathbf{X}, Y(C), Y(T))$, can be an important adjunct to propensity score methods, just as covariance adjustment can be an important adjunct in randomised experiments. Estimation can be improved when models are used to refine estimation, however it must be remembered that such modeling is a supplement to modeling the assignment mechanism, and is essentially adding a Bayesian component to the structure as in Rubin (1978).

2.5 Diagnostics for the Propensity Score: Goodness-of-Fit

Standardised differences for comparing means and prevalences between groups

Normand, Landrum, Guadagnoli, Ayanian, Ryad, Cleary and McNei (2001), Austin and Mamdani (2006), and Austin, Grootendorst and Anderson (2007) have proposed that standardised differences[16] be used to compare the mean of an observed baseline covariate between treated and untreated subjects in a propensity-score matched sample. The standardised difference (or bias) is defined as (Austin et al., 2007; Austin, 2008):

$$Bias = 100 \cdot \frac{\bar{x}_t - \bar{x}_c}{\sqrt{0.5 \cdot \left(s_t^2 + s_c^2\right)}} \tag{2.5.1}$$

for continuous variables, and as

$$Bias = 100 \cdot \frac{\hat{p}_t - \hat{p}_c}{\sqrt{0.5 \cdot \left(\hat{p}_t\left(1 - \hat{p}_t\right) + \hat{p}_c\left(1 - \hat{p}_c\right)\right)}} \tag{2.5.2}$$

[16] Standard difference is a convenient way to quantify the bias between treatment and control samples (Normand et al., 2001).

for dichotomous variables.

In formula (2.5.1), $\bar{x}_t$ and $\bar{x}_c$ denote the mean of continuous variable in treated and untreated or control subjects, respectively. The s_t^2 and s_c^2 denote the variances of the continuous variable in the treated and control subjects, respectively.

In formula (2.5.2), $\hat{p}_t$ and $\hat{p}_c$ denote the proportion of treated and untreated or control subjects, respectively, for whom the condition denoted by the dichotomous variables is present.

The standardised difference compares the difference in means in units of the pooled-standard deviation (Austin, 2008). Unlike t-tests and other statistical tests of hypothesis, the standardised difference is not influenced by sample size. Thus, the use of the standard difference can be used to compare balance in measured variables between treated and untreated subjects in the unweighted sample with that in the weighted sample (Austin, 2009). Furthermore, it allows for the comparison of the relative balance of variables measured in different units (e.g. age in years with systolic blood pressure in mm Hg). It has been suggested that a standardised difference of greater than 10% represents meaningful imbalance in a given variable between treatment groups (Normand et al., 2001).

A two-sample t-test is used to check if there are significant differences in covariate means for both groups. Before matching differences are expected, but after matching the covariates should be balanced in both groups and hence no significant differences should be found. This test might be preferred if the analyst is concerned with the statistical significance of the results (Caliendo, 2006). The shortcoming here is that the bias reduction before and after matching is not clearly visible (Caliendo, 2006).

The standardised difference provides a framework for comparing the mean or prevalence of a baseline covariate between treatment groups in the propensity score matched sample (Austin, 2009). However, a thorough examination of the comparability of treated and untreated subjects in the propensity score matched sample should not stop with a comparison of means and prevalences. The true propensity score is a balancing score: within strata matched on the true propensity score, the distribution of observed baseline covariates is independent of treatment status. Thus, the entire distribution of baseline covariates, not just means and prevalences, should be similar between treatment groups in the matched sample. Therefore, higher order moments of covariates and interactions between covariates should be compared between treatment groups (Austin, 2009).

Researchers may want to compare the distribution of a continuous variable between treated and untreated subjects in the matched sample. To accomplish this, researchers can use side-by-side boxplots, empirical cumulative distribution functions or nonparametric estimates of the probability density function. While standardised differences compare the difference in means between treated and untreated subjects, these graphical methods permit a broader comparison of the distribution of a continuous variable between two groups (Austin, 2008).

Stuart (2010) commented that matching methods are not themselves methods for estimating causal effects. Therefore, after the matching has created the comparable study groups, the researcher can move to the (outcome) analysis stage which may involve regression adjustments using the matched samples.

Chapter Summary

This chapter dealt with the issues regarding problem of causal inference under the counter-factual theory. The four formal modes of causal inference that are considered to be distinct were highlighted. Also, the discussion of sources of bias in observational studies and how matching is used to reduce it (bias) in the estimation of the intervention effect, and how it is introduced as a nonparametric method of adjustment for treatment assignment and as a method to form quasi-experimental contrasts was discussed.

A discussion of how matching is carried out in the absence of randomisation process (as in CRTs) was detailed. An overview of the methods to improve validity when randomisation is absent, which are regression analysis, stratification and matching techniques was given. These methods are often limited. We then introduced a balancing score called propensity score, the conditional probability of receiving treatment given pre-treatment covariates, together with its key feature that given a strong ignorable assumption, treatment assignment and the potential outcomes are independent. Finally, a discussion regarding how propensity score is estimated and evaluated was supplied.

Chapter 3

Clustered Observations

*****

> *The simplest argument, then, for multilevel modeling techniques is this: Because so much of what we study is multilevel in nature, we should use theories and analytic techniques that are also multilevel. If we do not do this, we can run into serious problems.*
>
> *Luke (2004)*

Chapter Preview

Chapter 3 looks at the models that are commonly used to predict and (or) compare groups when the outcome of interest is continuous (linear regression model), and when the outcome is binary (logistic regression model). Then a description of models that incorporate the correlateness of the observations is given i.e., the models for clustered data are studied. Mixed-effect models are discussed and the description of how these models are implemented in Stata is outlined.

The distinction between linear mixed and nonlinear mixed effects models is discussed. Then a discussion of the concepts of the unit of analysis and unit of observation given that in multilevel analysis observations can be at micro-level (level-1) or macro-level (level-2) is given. Finally, issues of spill-over in terms of the study design (here, crossover design), data contamination and (non)-adherence are addressed.

3.1 Introduction

At the heart of studies of causality is the notion of paradox[1]. In particular, researchers have to be more vigilant of Simpson's paradox (or Yule-Simpson effect). Formally, Yule-Simpson effect is not a paradox, because it does not lead to a contradiction. Nothing says in probability theory a statistical relationship cannot be reversed. Among several ways - this so-called paradox is addressed by fiiting a model that account for significant covariates.

One of the most important issues in statistical science is the construction of probabilistic models that represent, or sufficiently approximate, the true generating mechanism of a phenomenon under study (Larson, 1982; Ntzoufras, 2009). Usually models are constructed in order to assess or interpret causal relationship between response variable y and various characteristics expressed as variables called covariates or explanatory variables[2] (Ntzoufras, 2009).

Regression is the study of dependence (i.e. the process of finding the function satisfied by the points on the scatter diagram; and regression analysis[3], a statistical technique for investigating and modeling the relationship between variables, is a central part of many research projects (Montgomery and Peck, 1982; Weisberg, 2005). In another words, regression is a statistical method by which one variable is explained or understood on the basis of one or more other variables (Hilbe, 2009). As such, regression analysis[4] is a central part of many research projects.

The variable that is being explained is called the outcome, dependent or response, variable; the other variables used to explain or predict the response are called regressors, independent variables, covariates or predictors (Hilbe, 2009; Ntzoufras,

[1] A paradox is something that on the surface seems contradictory. Paradoxes help to reveal underlying truth beneath the surface of what appears to be absurd. For example, suppose we are observing several groups, and establish a relationship or correlation for each of these groups. Simpson's paradox says that when we combine all of the groups together, and look at the data in aggregate form, the correlation that we noticed before may reverse itself. This is most often due to lurking variables that have not been considered, but sometimes it is due to the numerical values of the data (http://statistics.about.com/od/HelpandTutorials/a/What-Is-Simpsons-Paradox.htm).

[2] Explanatory or predictor variable: A variable which is used in a relationship to explain or to predict changes in the values of another variable is called the dependent variable

[3] Formal assumptions of regression analysis are:

a) The errors all have expected value zero; $E(\epsilon_i)$ =0 for all i.

b) The errors all have the same variance; $Var(\epsilon_i) = \sigma_i^2$ for all i.

c) The errors are independent of each other.

d) The errors are all normally distributed; ϵ_i is normally distributed for all i.

[4] Regression analysis is used when two or more variables are thought to be systematically connected by a linear relationship.

2009).

The important instance of regression methodology is called linear regression (Nathans, Oswald and Nimon, 2012), and this method is the most commonly used in regression, and virtually all other regression methods build upon an understanding of how linear regression works (Hosmer and Lemeshow, 2000; Weisberg, 2005).

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data (Nathans et al., 2012). Before attempting to fit a linear model to observed data, one should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other, but that there is some significant association between the two variables.

A scatter plot can be a helpful tool in determining the strength of the relationship between two variables. Weisberg (2005) gives full account with regards to scatter diagrams. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model (Weisberg, 2005).

In a regression problem wherein only one predictor variable, generically called X and one response variable called Y, the data consist of values (x_i, y_i) where $i = 1, \cdots, n$, of (X, Y) observed on each of n units or cases. The goal of regression is to understand how the values of Y change as X is varied over its range of possible values. A first look at how X is varied is available from a scatter diagram.

3.1.1 The Mean Function and Variance Function

One important distribution of how Y changes as X is varied is the mean function, which is defined by $E(Y|X = x)$, and this is read as the expected value of the response when the predictor is fixed at the value of $X = x$; and this function depends on the values of x. The mean function depends on the problem, but generally the mean function is of the form

$$E(Y|X = x) = \beta_0 + \beta_1 x \tag{3.1.1}$$

This particular mean function is a (straight) line and has two parameters, an intercept β_0 and a slope β_1, which is the rate of change in $E(Y|X = x)$ for a unit change in X. The mean function would be completely specified if all β_i were known. Usually these constants need to be estimated from the data.

Another characteristic of the distribution of the response given the predictor is the variance function, see Fox (1997), defined as the variance of response distribution given that the predictor is fixed at $X = x$; symbolically written as

$$Var(Y|X = x). \tag{3.1.2}$$

A frequent assumption in fitting linear regression model is that the variance function is the same for every value of x. This is usually written as

$$Var(Y|X = x) = \sigma^2 \tag{3.1.3}$$

where σ^2 is a generally unknown constant (Fox, 1997; Weisberg, 2005).

3.1.2 Linear Regression: Model Interpretation

3.1.2.1 Simple Linear Regression

The simple linear regression consists of the mean function

$$E(Y|X = x) = \beta_0 + \beta_1 x \tag{3.1.4}$$

and the variance function

$$Var(Y|X = x) = \sigma^2 \tag{3.1.5}$$

Since the variance $\sigma^2 > 0$, the observed value of the i^{th} response will typically not equal its expected value $E(Y|X = x_i)$. To account for this difference between the observed data and expected value, a quantity called a statistical error (ϵ_i) for case i is used; and this quantity is defined implicitly as (Weisberg, 2005):

$$y_i = E(Y|X = x_i) + \epsilon_i \tag{3.1.6}$$

or explicitly as

$$\epsilon_i = y_i - E(Y|X = x_i) \tag{3.1.7}$$

The errors are unobservable quantities, and are random variables, as they depend on unknown parameters. Therefore, the errors correspond to the vertical distance between the point y_i and the mean function $E(Y|X = x_i)$. If the assumed mean function is incorrect, then the difference between the observed data and the incorrect mean function will have a nonrandom component (Fox, 1997).

There are two (important) assumptions about the errors. First, it is assumed that $E(\epsilon_i|x_i) = 0$, so a scatterplot of the ϵ_i versus the x_i would be null, with no patterns.

The second assumption is that the errors are all independent. Errors are often assumed to be normally distributed.

3.1.2.2 Multiple linear regression

Multiple linear regression generalises the simple linear regression model by allowing for many terms in the mean function rather than one intercept and one slope. Suppose the mean function of a simple linear regression is given. That is, we start with $E(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1$; and we wish to add the second variable X_2 with which to predict the response. The mean function that depends on both the value of X_1 and the value of X_2 is given by

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{3.1.8}$$

Weisberg (2005) commented that the inclusion of X_2 in a model attempts to explain the part that has not already been explained by X_1. Strickly speaking, the objective of multiple regression analysis is to use the covariates whose values are known to predict the single dependent value selected by the researcher or analyst (Hair, Anderson, Tatham and Black, 1998).

The normal (multivariate) linear model (Larson, 1982; Weisberg, 2005; Alexopoulos, 2010),

$$\begin{aligned} y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i, \\ &= \epsilon_i + \sum_{j=1}^{p} \beta_j x_{ji}, \\ \epsilon_i &\sim NID(0, \sigma^2) \end{aligned} \tag{3.1.9}$$

or simply

$$E(Y|X) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \tag{3.1.10}$$

has one random effect, the error term ϵ_i. The parameters of the model are the regression coefficients, $\beta_1, \beta_2, \cdots, \beta_p$, and the error variance, σ^2. Usually, $x_{1i} = 1$ [see equation (3.1.9)], and so β_1 is a constant or intercept.

The symbol X in $E(Y|X)$ (3.1.10) means that we are conditioning on all terms on the right side of the equation. Similarly, when we are conditioning on specific values for the predictors $x_1, \cdots, x_p$ that are collectively called, we write:

$$E(Y|X = x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{3.1.11}$$

As in simple linear regression, the β_i's are unknown parameters that need to be estimated. In particular, the regression coefficient represents the amount of change in the dependent variable for a one-unit change in the independent variable. In the multiple predictor model such as $Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$, the regression coefficients are partial coefficients because each takes into account not only the relationship between Y and $X_i, i = 1, \cdots, p$, but also between X_i and X_j for all $i \neq j$ (Hair et al., 1998, Page 149). The linear model is usually written in matrix form as,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon, \\ \epsilon &\sim \mathbf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n) \end{aligned} \tag{3.1.12}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ is the response vector; $\mathbf{X}$ is the model matrix, with typical row $\mathbf{x}_i^{'} = (x_{1i}, x_{2i}, \cdots, x_{pi})$; $\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$ is the vector of regression coefficients; $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$ is the vector of errors; $\mathbf{N}_n$ represents the n-variable multivariate-normal distribution; $\mathbf{0}$ is an $n \times 1$ vector of zeroes; and $\mathbf{I}_n$ is the order-n identity matrix.

3.1.3 Logistic Regression: Model Interpretation

Many educational research problems call for the analysis and prediction of a dichotomous outcome: whether a student will succeed in school, whether a child should be classified as learning disabled (LD), whether a teenager is prone to engage in risky behaviours, voter opinions ("Yes" vs "No"), test results ("Pass" vs "Fail") and so on.

Logistic regression[5] sometimes called the logistic model or logit model, analyses the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve (Hosmer and Lemeshow, 2000; Agresti, 2002; Hilbe, 2009).

There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed (Agresti, 1996; Hilbe, 2009).

[5] In logistic regression, instead of predicting the value of the dependent variable Y from a predictor variable X_i or several predictor variables (Xs), we predict the probability of Y occurring given known values of X (or Xs).

3.1.3.1 Simple logistic regression

Simple logistic regression analysis refers to the regression application with one dichotomous outcome (or one binary response) and one independent variable. For a binary response Y and a quantitative explanatory variable X, let

$$\begin{aligned}\pi(x) &= P(Y = 1|X = x),\\ &= 1 - P(Y = 0|X = x)\end{aligned} \tag{3.1.13}$$

denote the "success" probability when X takes values x. This probability is the parameter for the binomial distribution. The logistic regression model has linear form for the logit of this probability,

$$logit[\pi(x)] = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x. \tag{3.1.14}$$

Here, $logit[\pi(x)] = g(x)$. This formula (3.1.14) implies that $\pi(x)$ increases or decreases as an S-shaped function of x, see Figure 3.1, (Agresti, 1996; Hosmer and Lemeshow, 2000; Agresti, 2002). An alternative formula for logistic regression refers directly to the success probability and it is given as (Agresti, 1996, Page 103):

$$\pi(x) = \frac{e^{\{\beta_0+\beta_1 x\}}}{1 + e^{\{\beta_0+\beta_1 x\}}}. \tag{3.1.15}$$

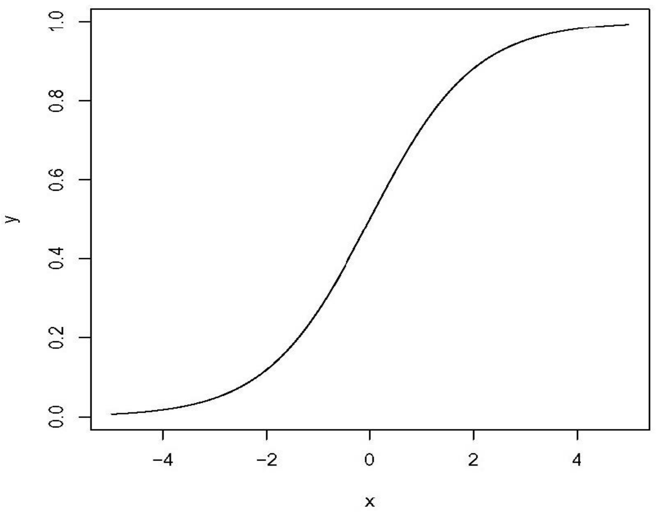


Figure 3.1: Relationship of a Binary Outcome, y (1 =Success, 0 = Failure) With a Continuous Predictor, x Scores

The parameter β_1 determines the rate of increase or decrease of the S-shaped curve. In particular, the sign of β_1 indicates whether the curve ascends or descents, and the rate of change increases as $|\beta_1|$ increases. When the model holds with $\beta_1 = 0$, the right-hand side of Equation (3.1.15) simplifies to a constant. Then, $\pi(x)$ is identical at all x. This implies that the binary outcome Y is independent of X (Agresti, 1996). In cases where $\beta_1 \neq 0$ then the steepest level of the curve occurs at *median effective level*[6] (EL_{50}), where $x = -\beta_0/\beta_1$. EL_{50} represents the level at which each outcome has 50% chance (Agresti, 1996).

If we assume that the independent variable is dichotomous such that

$$x = \begin{cases} 1, & \text{if success} \\ 0, & \text{otherwise} \end{cases},$$

then the difference in the logit (or logit difference) for subject with $x = 1$ and $x = 0$ is $g(1) - g(0) = \beta_1$.

The result β_1 is better interpreted through odds ratio (OR)[7]. Presenting the possible values of logistic probabilities in a 2×2 tabular format as shown in Table 3.1, we define the odds[8] of the outcome being realised among participants with $x = 1$ as:

$$\Omega_1 = \frac{\pi(1)}{1 - \pi(1)} \qquad (3.1.16)$$

Also, the odds of outcome being realised among participants with $x = 0$ is

$$\Omega_0 = \frac{\pi(0)}{1 - \pi(0)}. \qquad (3.1.17)$$

Odds are obtained by dividing the number of times an outcome of interest does happen by the number of times when it does not happen (Warner, 2013, Page 1013). The odds Ω are nonnegative, with $\Omega > 1.0$ when a success is more likely than a failure (Agresti, 2002, Page 44). In another words, the minimum value of odds is 0; this occurs when the frequency of the event in the numerator is zero (Warner, 2013).

The OR is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$, and is given by the expression (with proper substitution of entries of Table 3.1).

[6] For all values of $x = -\beta_0/\beta_1$, we have $\pi(x) = 0.5$.

[7] An odds ratio (OR) is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

[8] We note that for a probability π of the outcome of interest (e.g., success), the odds are defined to be $\Omega = \pi/(1 - \pi)$ (Agresti, 2002, Page 44).

$$\begin{aligned} OR &= \frac{\Omega_1}{\Omega_0} \\ &= \frac{\pi(1)}{1-\pi(1)} \times \frac{1-\pi(0)}{\pi(0)} \\ &= e^{\beta_1} \end{aligned} \tag{3.1.18}$$

Table 3.1: Values of the Logistic Regression Model for Dichotomous Covariate.
(Hosmer and Lemeshow, 2000, Page 49)

	Independent Variable (X)	
Outcome (Y)	$x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{e^{\{\beta_0+\beta_1\}}}{1+e^{\{\beta_0+\beta_1\}}}$	$\pi(0) = \frac{e^{\{\beta_0\}}}{1+e^{\{\beta_0\}}}$
$y = 0$	$1 - \pi(1) = \frac{1}{1+e^{\{\beta_0+\beta_1\}}}$	$1 - \pi(0) = \frac{1}{1+e^{\{\beta_0\}}}$

Table 3.2 summarises the possible interpretation of ORs. Odds ratios are most commonly used in case-control studies[9], however they are also being used in cross-sectional and cohort study designs as well (with some modifications and/or assumptions). ORs are used to compare the relative odds of the occurrence of the outcome of interest (e.g. disease or disorder), given exposure to the variable of interest (e.g. health characteristic, aspect of medical history). The OR can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.

An important part of most observational studies is the choice of control variables. In studying the effect of X on Y, one should control any covariate that can influence that relationship. When a non-casual association is observed between a given exposure and outcome is as a result of the influence of a third variable, it is termed confounding, with the third variable termed a confounding variable. A confounding variable is causally associated with the outcome of interest, and non-causally or causally associated with the exposure, but is not an intermediate variable in the causal pathway between exposure and outcome. Stratification and multiple regression techniques are two methods used to address confounding, and produce "adjusted" ORs (Lee, 2011). Among others, Lewallen and Courtright (1998), Hosmer and

[9] Case-Control study is a retrospective study that has been designed to help determine if an exposure is associated with an outcome (Lewallen and Courtright, 1998)

Lemeshow (2000), Agresti (2002) and Lee (2011) provide detailed explanation, and discussions with regard to interactions and confounding.

Table 3.2: Interpretation of OR: Values of OR farther from 1.0 in a given direction represent stronger association

Odds Ratio (OR)	Interpretation
Less than 1 $(OR < 1)$	Exposure associated with lower odds of outcome
Equals 1 $(OR = 1)$	Exposure does not affect odds of outcome
Greater than 1 $(OR > 1)$	Exposure associated with higher odds of outcome

3.1.3.2 Multiple logistic regression

The strength of a modeling technique lies in its ability to model many variables, some of which may be on different measurement scales (Hosmer and Lemeshow, 2000). A case in which more than one independent variables is generalised in logistic regression is referred to as the *multivariable case*. We consider a collection of p covariates denoted by $\mathbf{x}' = (x_1, x_2, \cdots, x_p)$.

If we define the conditional probability that the outcome is present as (Hosmer and Lemeshow, 2000):

$$P(Y = 1|\mathbf{x}) = \pi(\mathbf{x}) \tag{3.1.19}$$

then the logit of the multiple logistic regression model is given by the equation

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{3.1.20}$$

in which case the logistic regression model is

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}. \tag{3.1.21}$$

The parameter β_i refers to the effect of X_i on the log odds that $Y = 1$, controlling for other Xs. In another words, e^{β_i} is the multiplicative effect on the odds of a 1-unit increase in X_i, at fixed levels of the other Xs (Agresti, 1996, 2002).

A collection of design variables, also known as dummy variables is recommended for discrete and nominal independent variable, and as such discrete or categorical independent variables are not supposed to be treated as if they were continuous variable (Hosmer and Lemeshow, 2000, Page 32) and (Agresti, 2002, Page 183). Hosmer and

Lemeshow (2000) provide a very detailed description of logistic regression analysis and its applications.

3.2 Mixed Models

Mixed-effect models (or just mixed models) include additional random-effect terms, and are often appropriate for representing clustered data, dependent data or correlated data – arising, for example, when data are collected hierarchically (Murray, 1998).

In many applications, multiple measurements are made on the same experimental units (Bergsma, Croon and Hagenaars, 2009). In another words, in applied sciences, one is often confronted with the collection of correlated data (Abrahantes, Molenberghs, Burzykowski, Shkedy, Abad and Renard, 2004; Letsoalo and Lesaoana, 2010). This generic term embraces a multitude of data structures, such as multivariate observations, clustered data, repeated measurements, longitudinal data, and spatially correlated data (Abrahantes et al., 2004). When measurements are made over a period of time then such data are called repeated measures or clustered data.

The design for repeated measures could be one of the standard designs, e.g., a completely randomised design or a randomised complete block design. Correlated data originate in situations where observations in a sample are not selected independently of each other. This may happen in various settings, most of which are common to demographers and biomedical scientists. However, it is usually assumed that the repeated measurements from different study subjects (participants) are independent and correlated only when they come from the same subject (Wu and Zhang, 2006, Page 19).

3.2.1 Model Specification in Stata

The observations y_{ij} of participant j on occasion i can be modeled as

$$y_{ij} = \beta + \gamma_j + \epsilon_{ij} \tag{3.2.1}$$

where γ_j is the difference between the overall mean β and participant j's mean measurement (over hypothetical population of measurement occasions), and ϵ_{ij} is the measurement error for participant j on occasion i (Rabe-Hesketh, Skrondal and Pickles, 2002). The difference between the overall mean and the participant's mean γ_j has zero mean over participants and ϵ_{ij} has zero mean over occasions and participants. The model in (3.2.1) looks like a one-way ANOVA model with participant as a factor (categorical predictor). However, instead of fitting an ANOVA by estimating the γ_j as fixed effects, we assume that γ_j are random effects or random intercepts that are normally distributed (Rabe-Hesketh et al., 2002; Skrondal and Rabe-Hesketh, 2004),

$$\gamma_j \sim N(0, \Phi)$$

and independent of the ϵ_{ij} which are also specified as normal:

$$\epsilon_{ij} \sim N(0, \theta).$$

The ϵ_{ij} represent the effects of the occasions nested in particpants, as well as any other error (Skrondal and Rabe-Hesketh, 2004; Rabe-Hesketh, 2005).

3.2.2 Estimation using Stata's *xtreg*

The parameters of the variance-components model (3.2.1) can be estimated using *xtreg* command with *mle* option. Here, *mle* stands for 'maximum likelihood estimation' (Rabe-Hesketh, 2005, Page 9).

In Stata's commands for regression models the first variable name after the command name is the response variable. In *xtreg*, the fixed part of the model is specified next. For variance-components models, the fixed part is just intercept β, but this is included by default, so we need not specify any variables. The random part includes a random intercept γ_j whose cluster identifier is specified in the *i( )* option. The level-1 residual ϵ_{ij} need not be specified because it is always included (Skrondal and Rabe-Hesketh, 2003; Rabe-Hesketh, 2005). Here, the command therefore is

- *xtreg marks province, i(candidate) mle*.

We note that *xtreg* is applicable when dependent variable is a continuous variable. In case of a binary outcome - *xtlogit* with *pa* (population average) option is used (Skrondal and Rabe-Hesketh, 2003).

3.3 Multilevel Models

Many kinds of data, including observational data collected in the human and biological sciences, have hierarchical, nested, or clustered structure (Sullivan, Dukes and Losina, 1999; Ukoumunne, Gulliford and Chinn, 2004; Letsoalo and Lesaoana, 2010; Goldstein, 2011; Letsoalo and Lesaoana, 2012). Figure 3.2 presents a hypothetical example of clustered data.

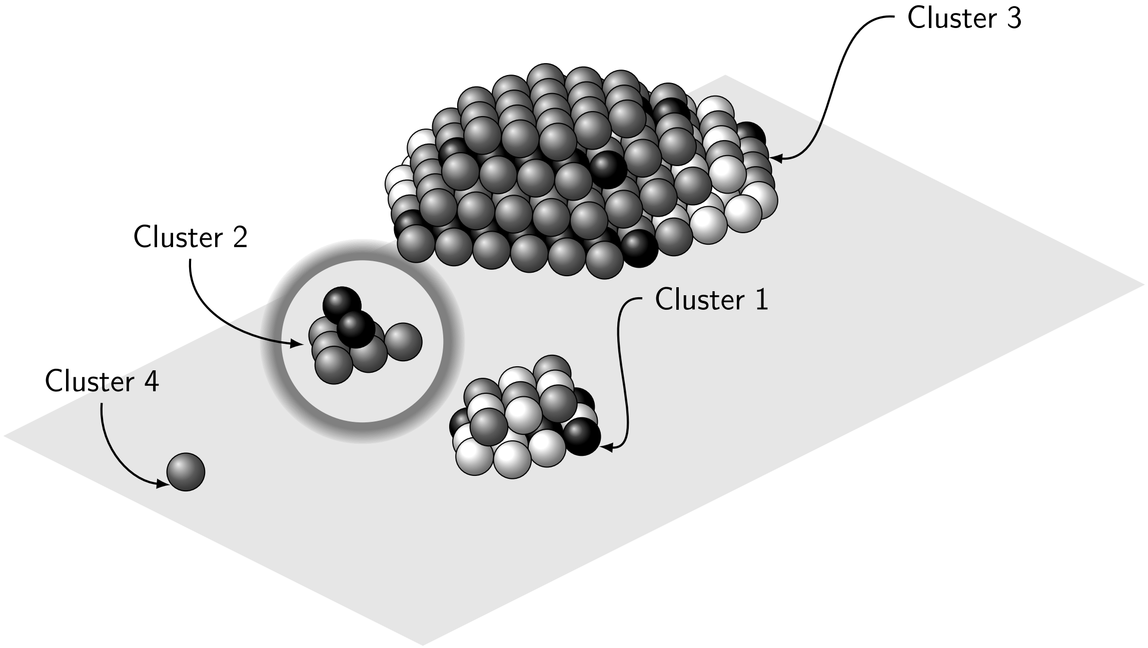
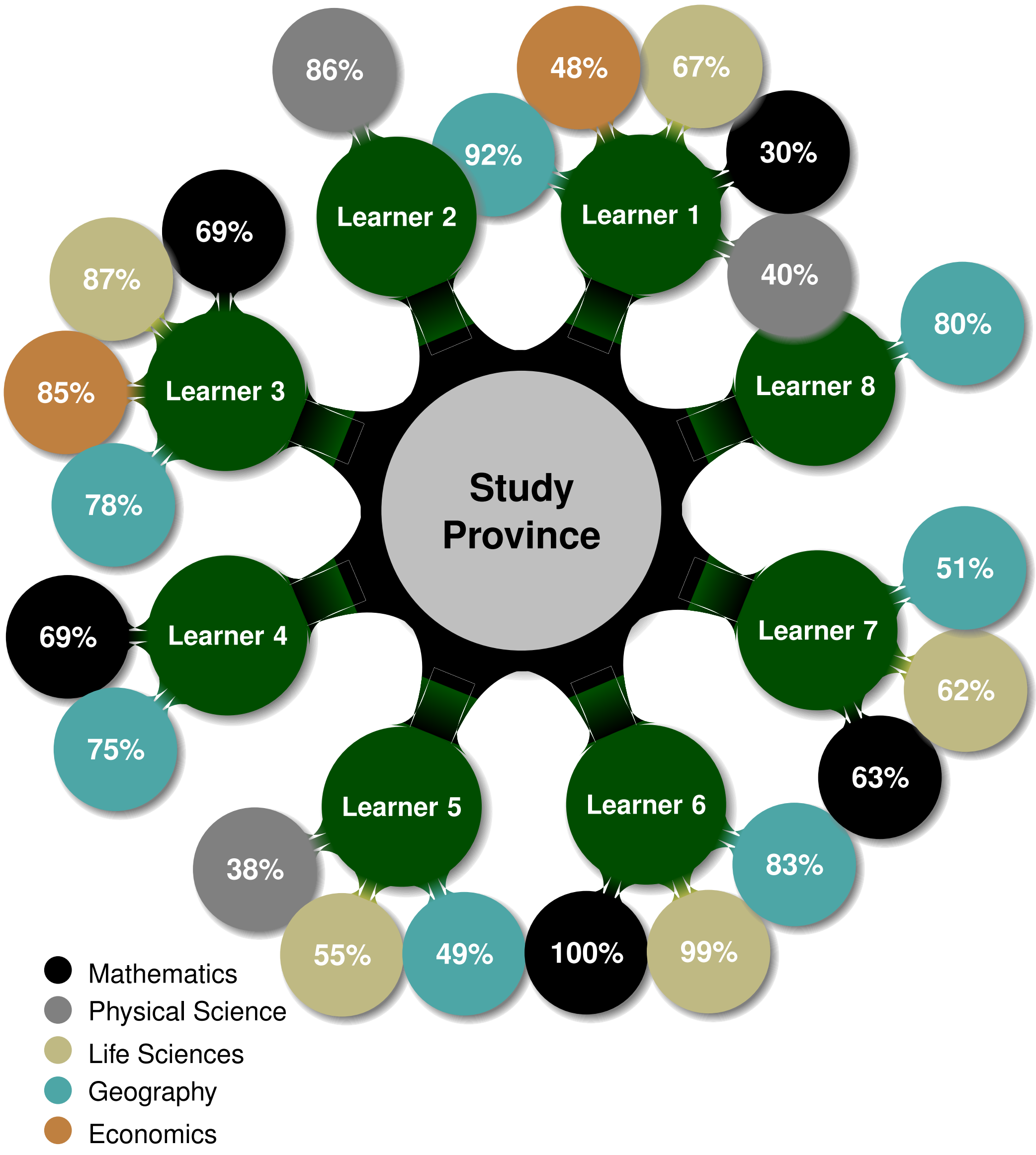


Figure 3.2: Hypothetical/example of clustered data
Note: Clusters need not be of equal sizes.
A cluster may represent a learner while atoms/balls represent subjects enrolled. Figure 3.3 gives a typical example of clustered data.
Source: http://www.texample.net/tikz/examples/clusters-of-atoms/

A hierarchy consists of units grouped at different levels (Ukoumunne et al., 2004). Schooling systems present an obvious example of hierarchical structure, with learners clustered within schools, which themselves may be clustered within education authorities. Learners may be the level-1 units clustered or nested within schools that are the level-2 units (Goldstein, 2011, Page 3). Figure 3.3 attempts to explain or presents a hypothetical example of clustered data in education setting. The goal of a multilevel model is to predict values of some dependent variable based on a function of predictor variables at more than one level (Luke, 2004).

Multilevel models (also called hierarchical linear models, nested models, mixed models, random coefficient, random-effects models, random parameter models or split-plot designs) are statistical models of parameters that vary at more than one level. These models can be seen as generalisations of linear models (in particular, linear regression), although they can also extend to non-linear models. Multilevel modeling has been developed as a technique for analysing data arranged in a variety of hierarchies of clusters or groups (Hox, 1995; Langford, Bentham and McDonald, 1998), that is, multilevel analysis is a methodology for the analysis of data with complex patterns of variability, with a focus on nested sources of variability (Langford et al., 1998; Snijders and Bosker, 1999). Therefore, a widely used approach to handling dependencies in the data is by means of random coefficient models

Figure 3.3: A hypothetical clustered data in education setting.
Learner performances in different learning areas.



(Bergsma et al., 2009).

Multilevel analysis is a stream that has two tributaries (Snijders and Bosker, 1999, Page 1):

(a) *Contextual analysis*[10], which was developed in social sciences, focused on the effect of the social context on individual behaviour.

[10] Contextual analysis is an analytical approach originally used in sociology to investigate the effect of collective or group characteristics on individual level outcomes. In contextual analysis, group level predictors (often constructed by aggregating the characteristics of individuals within groups) are included together with individual level variables in standard regressions with individuals as the units of analysis (contextual effects models). This approach permits the simultaneous examination of how individual level and group level variables are related to individual level outcomes (Diez Roux, 2002).

According to Smith (2011), human behaviour can be conceptualised as being influenced by three factors, namely

(i) a person's prior personal disposition;

(ii) the impingement of social environment on that person; and

(iii) the interaction between the predisposing and environmental factors.

Accordingly, these factors imply a multilevel analysis of at least two levels, level-1 and level-2. A contextual study exemplifies a multilevel analysis because it includes variables in individual (level-1) and on the environment (level-2) (Smith, 2011). Contextual effects are the cross-level interactions between the personal and environmental variables, and the study of these interactions defines *contextual analysis*.

(b) *Mixed effect models*[11], which are statistical models in the analysis of variance and in regression analysis where it is assumed that some of the coefficients are fixed and others are random.

According to Snijders and Bosker (1999) contextual modeling until about 1980 focused on the definition of appropriate variables to be used in ordinary least squares regression analysis. The main focus in the development of statistical procedures for mixed models up to the1980s was on random effects (i.e., random differences between classes in some classification system) rather than on random coefficients (i.e., random effects of numerical variables). The two streams, mixed models and contextual modeling came together to form multilevel analysis. Figure 3.4 attempts to make this explanation more explicit.

In investigating the relationship between the study participants and society, generally individuals interact with the social contexts to which they belong, meaning that individual persons are influenced by the social groups to which they belong, and that the properties of those groups are in turn influenced by the individuals who make up that group (Hox, 1995).

In biomedical setting - It is intuitive that people from the same area may be more similar to each other in relation to their health status than to people from other areas. In other words, persons with similar characteristics may have different degrees of health according to whether they live in one area or another because of differing

[11] The term mixed model refers to the use of both fixed and random effects in the same analysis. Fixed effects have levels that are of primary interest and would be used again if the experiment were repeated. Random effects have levels that are not of primary interest, but rather are thought of as a random selection from a much larger set of levels. Subject effects are almost always random effects, while treatment levels are almost always fixed effects.

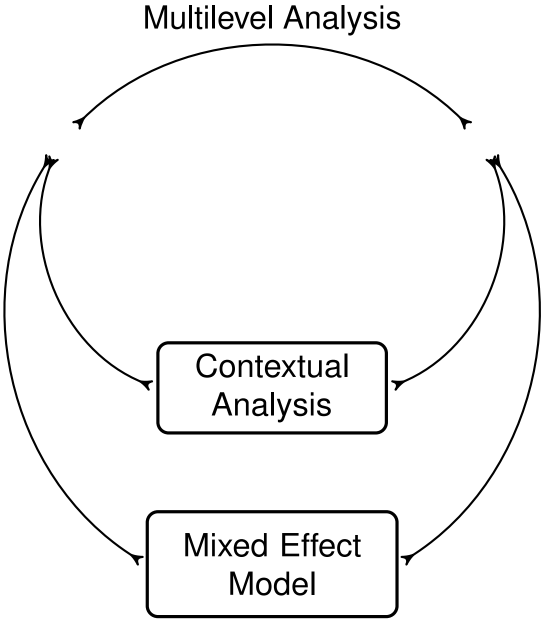


Figure 3.4: Tributaries of multilevel analysis
Source: Letsoalo (2004, Page 34)

cultural, economic, political, climatic, historical, or geographical contexts. Similarly, in educational setting - data are often organised at learner, classroom, school, and school district levels. This contextual phenomenon expresses itself as clustering of individual health status within areas. It follows that nested data or hierarchical data present problems for analysis.

People or creatures that exist within hierarchies tend to be more similar to each other than people randomly sampled from the entire population. Similarly, repeated measure observations on individual study participant (*as an example of clustered data*) tend to be correlated. In another words, the distinguishing feature of hierarchical or grouped data is that observations within a cluster may be correlated, and the degree of similarity among responses within a cluster is measured by a parameter called intracluster or intraclass correlation coefficient (ICC) (Donner, Piaggio and Villar, 2003; Letsoalo and Lesaoana, 2010).

The ICC may be interpreted as the standard Pearson correlation coefficient between any two responses in the same cluster (Donner et al., 2003). If the assumption that the ICC cannot be negative is added, then ICC may also be interpreted as the proportion of overall variation in response that can be accounted for by the between-cluster variation (Donner and Klar, 2000; Letsoalo and Lesaoana, 2010). A positive ICC implies that the variation between observations in different clusters exceeds the variation within clusters, hence it can be claimed that the design is characterised by 'between-cluster variation' (Donner et al., 2003; Donner and Klar, 2000).

In multilevel research, variables can be defined at any level of the hierarchy, and some of these variables may be measured directly at their natural level (Hox, 1995). In these schemes, individuals usually form micro-level, the lowest level or level-1, whereas groups (or clusters) form macro-level or level-2 (Letsoalo and Lesaoana, 2010).

A more general way to look at multilevel data is to investigate a cross-level hypothesis, or multilevel problem (Hox, 1995). A multilevel problem concerns the relationship between variables that are measured at a number of different hierarchical levels. Thus, multilevel models are designed to analyse variables from different levels simultaneously, using a statistical model that includes the various dependencies. Multilevel statistical models are always needed if a multi-stage sampling design has been employed (Levy and Lemeshow, 1999; Snijders and Bosker, 1999).

The main statistical model of multilevel analysis is the hierarchical linear model, an extension of the multiple linear regression model to a model that includes nested random coefficients. The multilevel regression model is also known as random coefficient model, mixed linear model, nested model or variance component model, due to the fact that these models have historically been used in educational research where hierarchies occur naturally (Hox, 1995; Sullivan et al., 1999; Letsoalo and Lesaoana, 2010).

3.4 Modeling Clustered Data: Linear Mixed-Effects and Nonlinear Mixed-Effects Models

In clustered data analysis and longitudinal studies, data from individuals are collected repeatedly over time whereas cross-sectional studies[12] only obtain one data point from each individual subject. Therefore, the key difference between clustered and cross-sectional data is that clustered data are correlated within a study participants (study subject) and independent between participants, while cross-sectional data are often independent (Wu and Zhang, 2006). Figure 3.3 on Page 77 presents a typical example of clustered data; where a learner was observed at least once.

A challenge for clustered data analysis is how to account for within-subject correlations. The results obtained from a standard statistical analysis which assumes all observations to be independent in clustered data, may be misleading. Such analysis is referred to as naïve pooling (Burton, Gurin and Sly, 1998). Therefore, standard

[12] Cross-sectional studies (also known as cross-sectional analyses, transversal studies, prevalence studies) are one type of observational study that involve data collection from a population, or a representative subset, at one specific point in time.

analysis, which ignores an important correlation structure, may well be misleading. One way of solving this problem is to create a single summary statistic such as the mean, for each cluster. This approach, referred to as data resolution, automatically avoids any over-inflation in the apparent size of the dataset (Burton et al., 1998; Letsoalo and Lesaoana, 2010).

Parametric mixed-effects models or random-effects models are powerful tools for clustered data analysis (Vaida, Meng and Xu, 2004); and linear and nonlinear mixed-effects models (including generalised linear and nonlinear mixed-effects models) have been widely used in many longitudinal studies[13] (Wu and Zhang, 2006, Page 17).

In particular, linear mixed-effects (LMEs) and nonlinear mixed-effects (NLMEs) models[14] are powerful tools for handling situations where one has to account for ICC when proper parametric models are available to relate a longitudinal or clustered response variable to its covariates (Vaida et al., 2004; Wu and Zhang, 2006, Page 2).

3.4.1 Linear Mixed-Effects Models

Linear mixed-effect models are used when the relationship between a clustered response variable and its covariates can be expressed via a linear model. LME models are also known as multilevel models, linear mixed-effects models, random-effects models, random-coefficient models, or hierarchical linear models. LME models handle unequal n_i's, time-varying covariates, and unequally spaced responses. LME models were first proposed as (for example, see Laird and Ware (1982)):

$$\begin{aligned} &y_{ij} = \mathbf{x}_{ij}^T \beta + \mathbf{z}_{ij}^T \mathbf{b}_i + \epsilon_{ij}, \\ &\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \epsilon_i \sim N(\mathbf{0}, \mathbf{R}_i), \\ &\ j = 1, 2, \cdots, n_i; i = 1, 2, \cdots, n, \end{aligned} \tag{3.4.1}$$

where $\epsilon_i = [\epsilon_{i1}, \cdots, \epsilon_{ini}]^T, y_{ij}$ and ϵ_{ij} denote the response and the measurement error of the j^{th} measurement of the i^{th} subject, the unknown parameters $\beta : p \times 1$ and $\mathbf{b}_i$ are usually called the fixed-effects vector and random-effects vectors, respectively, and $\mathbf{x}$ and z are the associated fixed-effects and random-effects covariate vectors, and $\mathbf{D}$ and $\mathbf{R}_i, i = 1, 2, \cdots, n$ are variance components of the LME model (Wu and Zhang, 2006, Page 18).

[13] A longitudinal survey is a correlational research study that involves repeated observations of the same variables over long periods of time - often many decades. It is a type of observational study.

[14] A mixed model is a statistical model containing both fixed effects and random effects, that is mixed effects.

The LME model (3.4.1) can be generally written as

$$\begin{aligned} &\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, \\ &\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \epsilon_i \sim N(\mathbf{0}, \mathbf{R}_i), \\ &\ i = 1, 2, \cdots, n, \end{aligned} \tag{3.4.2}$$

where $\mathbf{y}_i$ and ϵ_i are, respectively, the vectors of responses and measurement errors for the i^{th} subject, β and $\mathbf{b}_i$ are respectively, the vectors of fixed-effects (population-parameters) and random-effects (individual parameters), and $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the associated fixed-effects and random-effects design matrices, respectively (Wu and Zhang, 2006; Goldstein, 2011).

The mean and covariance matrix of $\mathbf{y}_i$ is given by:

$$\begin{aligned} E(\mathbf{y}_i) &= \mathbf{X}_i\beta, \\ Cov(\mathbf{y}_i) &= \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{\mathbf{T}} + \mathbf{R}_i, i = 1, 2, \cdots, n. \end{aligned} \tag{3.4.3}$$

In matrix notation (or compactly), the general LME model (3.4.2) can be further written as:

$$\begin{aligned} &\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon, \\ &\mathbf{b} \sim N(\mathbf{0}, \widetilde{\mathbf{D}}), \epsilon_i \sim N(\mathbf{0}, \mathbf{R}_i), \\ &i = 1, 2, \cdots, n, \end{aligned} \tag{3.4.4}$$

where

$$\begin{aligned} \mathbf{y} &= [\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T]^T, \\ b &= [\mathbf{b}_1^T, \cdots, \mathbf{b}_n^T]^T, \\ \epsilon &= [\epsilon_1^T, \cdots, \epsilon_n^T]^T, \\ \mathbf{X} &= [\mathbf{X}_1^T, \cdots, \mathbf{X}_n^T]^T, \\ \mathbf{Z} &= diag(\mathbf{Z}_1, \cdots, \mathbf{Z}_n), \\ \mathbf{D} &= diag(\mathbf{D}_1, \cdots, \mathbf{D}_n), \\ \mathbf{R} &= diag(\mathbf{R}_1, \cdots, \mathbf{R}_n). \end{aligned} \tag{3.4.5}$$

Based on the general LME model (3.4.4), we have $Cov(\mathbf{y}) = diag(Cov(\mathbf{y}_1), \cdots, Cov(\mathbf{y}_n))$ and the covariance matrix $Cov(\mathbf{y}_i)$ for repeated or clustered measurement vector $\mathbf{y}_i$ for the i^{th} subject is given in Equation (3.4.3).

Wu and Zhang (2006) state that the inference for β and $\mathbf{b}_i, i = 1, 2, \cdots, n$ for the general LME model (3.4.2) can be based on the likelihood method or generalised least square method. For known $\mathbf{D}$ and $\mathbf{R}_i, i = 1, 2, \cdots, n$, the estimates of β and $\mathbf{b}_i, i = 1, 2, \cdots, n$ may be obtained by minimising the following twice negative logarithm of the joint density function of $\mathbf{y}_i$ and $\mathbf{b}_i, i = 1, 2, \cdots, n$ (up to a constant):

$$\begin{aligned} GLL(\beta, \mathbf{b}_i|\mathbf{y}_i) = \sum_{i=1}^{n}\{[\mathbf{y}_i - \mathbf{X}_i\beta - \mathbf{Z}_i\mathbf{b}_i]^T\mathbf{R}_i^{-1}[\mathbf{y}_i - \mathbf{X}_i\beta - \mathbf{Z}_i\mathbf{b}_i] \\ + \mathbf{b}_i^T\mathbf{D}^{-1}\mathbf{b}_i + log|\mathbf{D}| + log|\mathbf{R}_i|\}. \end{aligned} \tag{3.4.6}$$

But $\mathbf{b}_i, i = 1, 2, \cdots, n$ are random-effects parameter vectors, the expression (3.4.6) is not a conventional log-likelihood (Wu and Zhang, 2006), however, (3.4.6) is called a generalised log-likehood (GLL) of the mixed-effects parameters $(\beta, \mathbf{b}_i, i = 1, 2, \cdots, n)$.

For given $\mathbf{D}$ and $\mathbf{R}_i, i = 1, 2, \cdots, n$, minimising the GLL criterion (3.4.6) is equivalent to solving the so-called mixed model equations:

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \tilde{D}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix},$$

where $\mathbf{y}, \mathbf{b}, \mathbf{X}, \mathbf{Z}, \tilde{D}$ and $\mathbf{R}$ are defined in (3.4.5). The respective estimates of β and $\mathbf{b}_i$ are given by:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}^{-1})\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{3.4.7}$$

and

$$\hat{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\beta}), i = 1, 2, \cdots, n, \tag{3.4.8}$$

where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i, i = 1, 2, \cdots, n$ and $\mathbf{V} = diag(\mathbf{V}_1, \cdots, V_n)$. The covariance matrices of $\hat{\beta}$ and $\hat{\mathbf{b}}_i$ are, respectively:

$$Cov(\hat{\beta}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} = \left(\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{D}, \tag{3.4.9}$$

$$\begin{aligned} Cov(\hat{\mathbf{b}}_i - \mathbf{b}_i) = \mathbf{D} - \mathbf{D}\left(\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right)\mathbf{D} + \mathbf{D}\left(\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right) \\ \times \left(\sum_{j=1}^{n}\mathbf{X}_j^T\mathbf{V}_j^{-1}\mathbf{X}_j\right)^{-1}\left(\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{Z}_i\right)^{-1}\mathbf{D}, \\ i = 1, 2, \cdots, n. \end{aligned} \tag{3.4.10}$$

3.4.2 Nonlinear Mixed-Effects Models

Nonlinear mixed-effects models[15] are used when the relationship between a clustered response variable and its covariates can be expressed via a nonlinear model, which is known except for some parameters (Zhang, Lin and Sowers, 2007). The NLME models are fully parametric and model the within-subject covariance structure more explicitly (Azzimonti, Ieva and Paganoni, 2013). NLME model may be written as (Wu and Zhang, 2006; Zhang et al., 2007; Azzimonti et al., 2013):

$$\begin{aligned} &\mathbf{y}_i = \mathbf{f}(\mathbf{X}_i, \beta_i) + \epsilon_i, \beta_i = \mathbf{d}(\mathbf{A}_i, \mathbf{B}_i, \beta, \mathbf{b}_i) \\ &\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \\ &\epsilon_i \sim N(\mathbf{0}, \mathbf{R}_i), i = 1, 2, \cdots, n, \end{aligned} \tag{3.4.11}$$

where $\mathbf{f}(\mathbf{X}_i, \beta_i) = [\mathbf{f}(\mathbf{x}_{i1}, \beta_i), \cdots, \mathbf{f}(\mathbf{x}_{in_i}, \beta_i)]^T$ with $\mathbf{f}(\cdot)$ being unknown function, $\mathbf{X}_i = [\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i}]^T$ a design matrix and β_i a subject-specific parameter for the i^{th} subject. Here, $\mathbf{d}(\cdot)$ is a known function of the design matrices $\mathbf{A}_i$ and $\mathbf{B}_i$, the fixed-effects vector β and random-effects vector $\mathbf{b}_i$.

The successful application of a LME model or a NLME model to longitudinal data analysis or clustered data analysis strongly depends on the assumption of a proper linear or nonlinear model for the relationship between response variable and the covariates. In cases where the assumption does not hold - the relationship between the response variable and the covariates has to be modeled nonparametrically.

The clustered or longitudinal dataset can be expressed in the form (Wu and Zhang, 2002, 2006):

$$(t_{ij}, y_{ij}), j = 1, 2, \cdots, n_i; i = 1, 2, \cdots, n, \tag{3.4.12}$$

where t_{ij} denotes the design time points (occasions), y_{ij} the responses observed at t_{ij}, n_i the number of observations for the n^{th} subject, and n is the number of subjects.

For such dataset, we do not assume a parametric model for the relationship between the response variable and covariate occasion. Instead, we assume that the individual and population mean functions are smooth functions of time t, and let the data themselves determine the form of the underlying functions (Wu and Zhang, 2006).

Wu and Zhang (2002, 2006) introduced a nonparametric mixed-effects (NPME) model as:

$$y_i(t) = \eta(t) + v_i(t) + \epsilon_i(t), i = 1, 2, \cdots, n, \tag{3.4.13}$$

[15] Nonlinear mixed-effects NLME models are mixed-effects models in which at least one of the fixed or random effects appears nonlinearly in the model function (Azzimonti et al., 2013)

where $\eta(t)$ models the population mean function of the clustered or longitudinal dataset, called fixed-effect function, $v_i(t)$ models the departure of i^{th} individual function from the population mean function $\eta(t)$, called the i^{th} random-effect function, and $\epsilon_i(t)$ the measurement errors that cannot be explained by both the fixed-effect and the random-effect functions.

The NLMEs model is a natural generalisation from the LME model. Suppose (μ, Ω) denotes distribution with mean vector μ and covariance matrix Ω. Then a NLME model can be written into a two-stage hierarchical form as:

Stage 1. Intra-subject variation

$$\begin{aligned} \mathbf{y}_i &= f(\mathbf{X}_i, \beta_i) + \epsilon_i | \beta \\ &\sim (0, \mathbf{R}_i(\beta_i, \xi)), \\ i &= 1, 2, \cdots, n, \end{aligned} \tag{3.4.14}$$

where $\mathbf{y}_i$ and ϵ_i are $(n \times 1)$ vectors of the responses and measurement errors for subject i, respectively; the vector function $\mathbf{f}(\mathbf{X}_i, \beta_i) = [f(\mathbf{x}_{i1}, \beta_i), \cdots, f(\mathbf{x}_{in}, \beta_i)]^T$ where $f(\cdot)$ is a known function, $\beta_i : p \times 1$ is an unknown parameter vector, and the design matrix $\mathbf{X}_i = [x_{i1}, \cdots, x_{in}]^T$; and the covariance matrix $\mathbf{R}_i(\cdot)$ is a known function with ξ being an unknown parameter vector (Wu and Zhang, 2006; Zhang et al., 2007).

Stage 2. Inter-subject variation

$$\begin{aligned} \beta_i &= \mathbf{d}(\mathbf{a}_i, \beta, \mathbf{b}_i), \\ \mathbf{b}_i &\sim (\mathbf{0}, \mathbf{D}), \\ i &= 1, 2, \cdots, n, \end{aligned} \tag{3.4.15}$$

where $\mathbf{d}(\mathbf{a}_i, \beta, \mathbf{b}_i)$ is a known p-dimensional function of the between-subject covariate vector $\mathbf{a}_i$, with the population parameter vector (known also as fixed-effects parameter vector) β and the random-effects vector $\mathbf{b}_i$. The function $\mathbf{d}(\cdot)$ may be linear or nonlinear. Wu and Zhang (2006; Page 27) gives details with regard to two-stage methods.

3.5 The Unit of Analysis

One of the most important ideas and/or aspects in any research project is what is regarded as the major entity that a researcher is analysing, the unit of analysis, because this has an impact on the analysis to be carried out. As such, the phrase *unit of analysis* is the source of much confusion in the context of group-randomised trials (Murray, 1998).

Almost anything can be the unit of analysis in a social research, yet the branch of social science mostly delimit, albeit not severely, the possibilities of the unit of analysis of a scientific study. In educational research, the most common units of analysis are students, parents, teachers, classes, schools or school districts.

Murray (1998) indicates that part of the confusion arises from the fact that there are often so many units from which to choose that analysts may easily and unknowingly choose badly. It is sometimes not clear what is the appropriate level of analysis. For instance, persons are in groups (e.g., learners in classrooms), and either person or group could be the unit of analysis. Therefore, the group would be the unit of analysis by computing a mean of those persons who are members of the group. Another source of confusion is that there are many different ways to conduct the analysis (Murray, 1998; Silverman and Solmon, 1998).

Therefore, the unit of analysis is the basic entity or object about which generalisations are to be made based on an analysis, and for which data have been collected. The unit of analysis is fundamental to data and statistical inference, to statistical data structures, and to secondary data analysis. The unit of analysis is the level at which data are used to represent one data point in an analysis (Silverman and Solmon, 1998).

The unit of analysis can be artifacts, areas, groups, individuals or social interactions such as divorces, marriages, etc. In choosing the unit of analysis the researcher should be aware of the possibility of the following fallacies, or errors of reasoning based on mistaken assumptions:

(a) The *ecological fallacy* also called ‘Robinson effect’ occurs when one makes conclusions about individuals based only on the analyses of group data. That is, interpreting aggregated data at the individual level. The ecological fallacy arises because association between two variables at the group level (or ecological level) may differ from associations between analogous variables measured at the individual level (Hox, 1995; Diez Roux, 2002).

(b) An *exception fallacy* occurs when one reaches a cluster/group conclusion on the basis of exceptional cases.

(c) *Atomistic fallacy*, an error committed when one draws inference at a higher level from analyses performed at a lower level. In another words, this fallacy arises because association between two variables at the individual level may differ from association between analogous variables measured at the group level (Diez Roux, 2002).

The unit of inference in group randomisation trials or clustered data analysis may be directed either at the group level or at the level of the individual study participant, and an interest in group-level inferences leads investigators to collect data only at the group level (Donner and Klar, 2000). The investigator faced with analysing individual-level data must account for the lack of statistical independence among observations within a group. The proper unit of analysis is determined entirely by the design of the study, including the selection and allocation schemes for group and members (Murray, 1998, Page 105).

Donner and Klar (2000, Page 81) suggests as a method of simplifying the problem, *to collapse the data in each cluster, followed by the construction of a meaningful summary measure to serve as the unit of analysis so that the standard statistical methods can be directly applied to the collapsed measures*. This approach is called aggregated analysis. Challenges in identifying and distinguishing the unit of inference, the unit of analysis and the impact of clustering are not unique to cluster randomisation trials, since Donner and Klar (2000) reported that many of the ideas presented in cluster randomisation trials are based on discussions of methods for the analysis of longitudinal or repeated-measures data.

Group-level analyses can be used for more complex summary scores and, more generally, for any study outcome. These analyses (group-level) are most obviously appropriate when the primary questions of interest are more concerned with the randomised unit as a whole, than with the individual cases/subjects (Murray, 1998).

Donner and Klar (1994) emphasise the fact that the primary advantage of using the cluster as the unit of analysis is that standard techniques are generally applicable for any one of the three designs. There are, however, disadvantages to this approach. Tests of significance using the cluster as the unit of analysis will, in general, have less power than methods using the individual as the unit of analysis. That is, the statistical power of group-randomised trials is greatly reduced in comparison with similar sized individually randomised trial (Murray, 1998). However, Donner and Klar (1994) quoting Shirley and Hickling (1981) point out that simulation studies have demonstrated that the loss in power is small, and using weighted least squares could increase the efficiency of these procedures.

Questions concerning the appropriate unit of analysis are more challenging when the primary target of inference is at the level of the individual subjects, with the choice of randomisation unit then largely a matter of convenience or other practical considerations. Particular care must be taken when conducting individual-level analysis to properly adjust for the effect of clustering (Donner and Klar, 2000, Page 81). Using an inappropriate unit of analysis may lead to results that are erroneous (Silverman and Solmon, 1998). Luke (2004) emphasised that fallacies are a problem of inference, not of measurement. That is, it is perfectly admissible to characterise a higher-level collective using information obtained from lower-level members. The types of fallacies described in this thesis come about when relationships discovered at one particular level are inappropriately assumed to occur in the same fashion at some other (higher or lower) level (Silverman and Solmon, 1998; Luke, 2004).

Generally, the unit of analysis chosen has consequences for research design, the number of participants or number of classes needed, and the faith we can place in the results and conclusions (Murray, 1998; Silverman and Solmon, 1998). Several strategies can be used to obtain valid analyses of cluster randomised experiments. The simplest is to treat the clusters as units of analysis by computing the mean scores on the outcome (and all other variables that may be involved in the analysis) for each cluster (e.g., classroom or school) and carrying out the statistical analysis as if the cluster means were the data (see the definition of naïve-pooling). Abrahantes et al. (2004) discuss the choice of unit of analysis and modeling strategies in clustered data analysis.

3.6 Transmission of Information: Spill-Over Effects

Randomisation is sometimes used to limit the possibility of transmission of information or data. Thus, randomisation is widely used when individual allocation is possible but there is concern over spill-over of information. Depending on the nature of the study, transmission of data or information can occur between or within clusters (or study groups).

There are three related concepts which have to do with transmission of information provided in one study arm to participants in the other study arm:

a) **Contamination**

The Contamination[16] involves an intervention intended for one study arm being

[16] Contamination refers to a situation where at least some of the information provided in one arm of

actively received without official knowledge or intention (and often incompletely, or incorrectly) by some participants in the other arm (Hayes and Moulton, 2009). It mostly happens where an intervention is easily transferred (such as information or tablets), when participants do not lose anything by passing on their treatment, or where there exists social/physical proximity between participants. Figure (3.5) attempts to makes this explanation more explicit. Contamination cannot occur when the intervention is highly targeted, e.g. injections (Briggs, 2003).

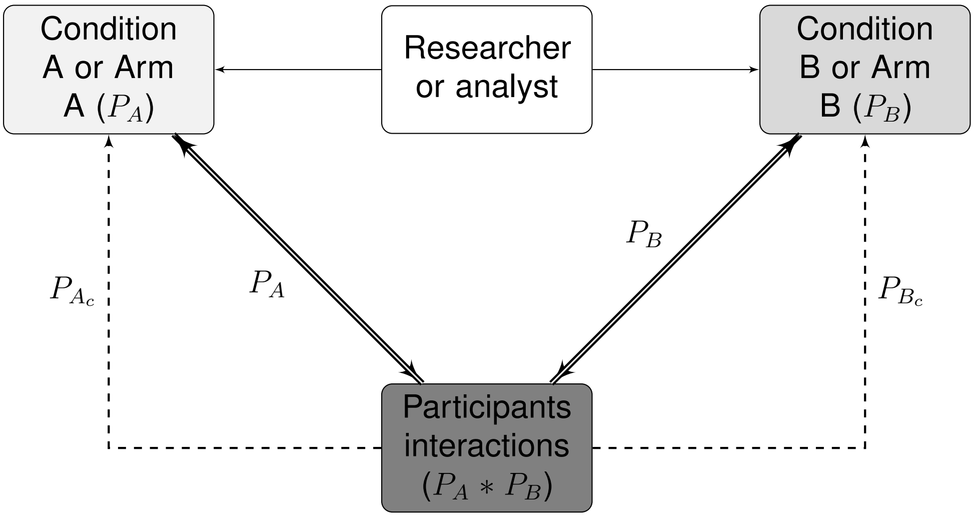

Condition A or Arm A (P_A)
Researcher or analyst
Condition B or Arm B (P_B)
P_{A_c}
P_A
P_B
P_{B_c}
Participants interactions ($P_A * P_B$)

Figure 3.5: Contamination occurs when individuals randomised or allocated to the study conditions, A or B, are exposed to the wrong condition through having contact with each other. Contamination can occur either inadvertently or intentionally as people discuss their experiences. The cost to internal validity is that people in the "control" condition receive part of the intervention.
Note P_{A_c} and P_{B_c} are contaminated arm A and contaminated arm B participants, respectively. Likewise, P_A and P_B are pure participants in arm A and arm B, respectively.

Another setting in which contamination is minimised is in group randomised controlled trials. These trials are attractive in settings in which individual randomisation is difficult or impossible (Isaakidid and Loannidid, 2003). Justification for adopting a group randomisation design rests on practical considerations such as the desire to control costs or attempting to minimise experimental contamination, and ethical considerations and should always be stated explicitly (Donner and Klar, 2000; Hayes and Moulton, 2009). Hayes and Moulton (2009) discuss ways in which contamination can occur in clinical randomised controlled trials.

b) Cross-Over

Cross-over is used to refer to a participant who is randomised to one trial arm, but

a trial is transmitted via informal pathways to participants in the other study arm.

receives (with official knowledge) the treatment from the other arm. In a cross-over trial, each participant gets both treatments being tested. A cross-over design is a modified, randomised block design in which each block, which may be a subject or a group of subjects, receives more than one formulation of a drug or treatment at different time periods. A cross-over design is called a complete cross-over design if each sequence contains each of the formulations (Chow and Liu, 2000).

Cross-over trials are trials in which patients are allocated to sequences of treatment with the purpose of studying differences between individual treatments (Sindrup, Andersen, Madsen, Smith, Brosen and Jensen, 1999). In another words, cross-over trials allow the response of a subject to treatment A to be contrasted with the same subject's response to treatment B. Sibbald and Roberts (1998) emphasise that removing patient variation in this way makes cross-over trials potentially more efficient than similar sized, parallel group trials in which each subject is exposed to only one treatment. In theory treatment effects can be estimated with greater precision given the same number of subjects (Jones and Kenward, 1989; Sibbald and Roberts, 1998).

Sindrup et al. (1999) give an example of a cross-over trial:

> ”*A cross-over trial was run to compare the effect of tramadol to placebo in painful polyneuropathy. Forty-five patients were randomised to one of two sequences, tramadol followed by placebo or placebo followed by tramadol. Each treatment was delivered for four weeks. Using 10-point numeric scales, patients rated pain, paresthesia and touch-evoked pain.*”

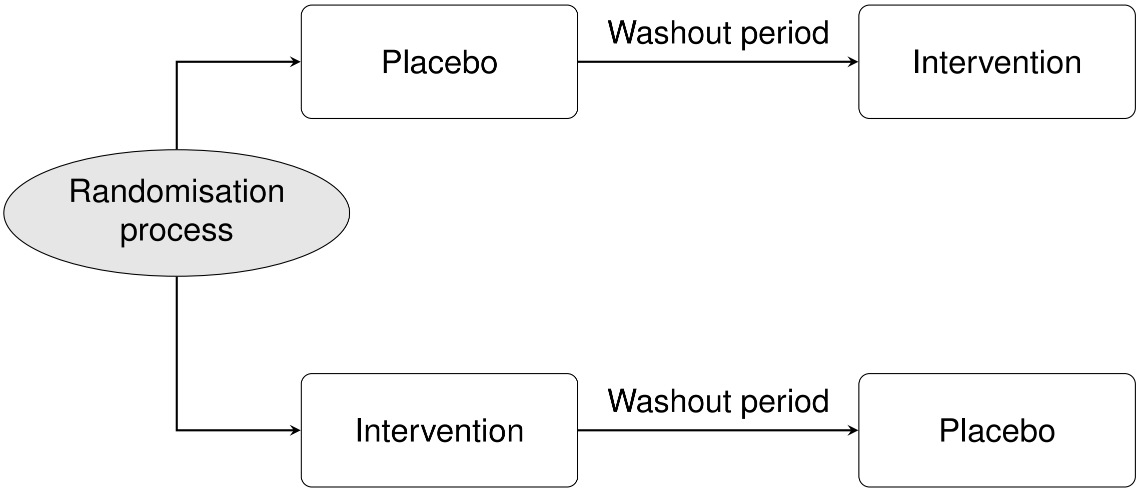


Figure 3.6: AB/BA Randomisation model

This is an example of the most common sort of cross-over, often loosely referred to as a two-period design, but more accurately described by referring explicitly to the sequences as an *AB/BA* design (Chow and Liu, 2000). In this case we have A as placebo and B as tramadol. Randomisation for a 2×2 cross-over design can be carried out by using either a table of random numbers or some procedures as embedded in some statistical software packages such as PROC PLAN in SAS.

To produce valid results, the effect of the first drug must end before the second drug is taken, and vice-versa; and this phenomenon is called *carry-over effect* (Stone, 1986; Jones and Kenward, 1989; Chow and Liu, 2000; Patterson and Jones, 2006). This requirement can be hard to satisfy, and is one reason cross-over trials are not often used. A washout period[17] between the two treatments might minimise the effects of the carry-over. Figure 3.6 attempts to make this explanation more explicit.

Carry-over effects can be avoided with a sufficiently long *wash-out* period between treatments. In another words, when an adequate wash-out period is included, carry-over effects are generally considered to be negligible (Patterson and Jones, 2006, Page 22). However, the planning for sufficiently long wash-out periods does require expert knowledge of the dynamics of the treatment, which often is unknown. Jones and Kenward (1989) and Chow and Liu (2000) give detailed discussions and statistical models for cross-over designs.

When carry-over effects are present, a standard 2×2 cross-over design may not be admirable, for it may not provide estimates for some fixed effects. To overcome such a challenge, a higher-order cross-over design, a design in which either the number of periods is greater than the number of formulations to be compared, or the number of sequences is greater than the number of formulations to be compared, may be useful (Chow and Liu, 2000). An in-depth discussion with regards the higher-order designs can be found in Jones and Kenward (1989).

Similar principle that is employed in psychology and social science studies is called counterbalancing, a method of controlling for order effects in a repeated measure design by either including all orders of treatment or randomly determining the order for each subject (Cozby, 2009). In essense, countebalancing principle or cross-over principle follows the within participants design or repeated measure design.

[17] The washout period is the rest period between two treatment periods for which the effect of one formulation administered at one treatment period does not carry over to the next (Chow and Liu, 2000, Page 38).

c) **Non-adherence**

Non-adherence (or non-compliance) refers to participants not fully receiving the treatment allocated, for instance, not taking all the medication as per prescription, not turning up for follow-up inspection, or not reading the information leaflet. Many factors are involved in participant non-adherence; factors related to the characteristics of the disease, medication side effects, duration of treatment, frequency of expected intake, complexity of treatment, and severity of the condition or disease. It has been demonstrated, for example, that people are less likely to continue their medication regimen over long periods and are less likely to be adherent when the daily doses increase from 1 pill to 4 pills (Kramer, 1995).

The problem of non-adherence to treatments, interventions or medications is serious, but not insurmountable. With each passing day, tremendous progress is being made to understand the core reasons for non-adherence and design programs that will address these issues. Also, there has been a realisation by all concerned stakeholders that they need to stop viewing non-adherence as either 'my problem' or 'their problem' and treat it as 'our problem' (Haynes, 2001).

The effects of these concepts or phenomena on the value of trial-data are similar – they bias the study results, tending to reduce the apparent size of any real difference in treatment effects, and increasing the chance of Type II error, also known as a false negative. The circumstances in which these phenomena are major problems are likely to be different, as are the measures to minimise them.

Learners tend to share study materials, for example; they can share study notes, or text books, also they tend to interact in study groups to share their learning experience. It is in these kinds of interactions that information from one school is filtered into another school especially when schools are under different authorities and/or are in different provinces, hence contamination or spill-over effect. For example, an intervention might target only *poor* children (the target group) within a locality (the local setting). In many cases, the local nontarget population may also be indirectly affected by the treatment through social and economic interaction with the treated individuals. These possible *interactions* are what we define as *spill-over effects*.

Also, nonadherence biases the results when learners do not comply with rules, conditions, or instructions from educators. Nonadherence occurs when learners do not hand in homeworks and/or assignments, and at times it happens when learners start bunking off classes or school.

Chapter Summary

Field experiments often assign entire intact groups (such as regions, classrooms or schools) to the same treatment group, with different intact groups assigned to different treatment. A common mistake in analysis of such data (clustered data) is to ignore the effect of clustering and analyse the data as if each study arm were a simpe random sample or observations were independent; for this may lead to an overstatement of the precision and anticonservative conclusions about the precision and statistical significance of intervention effects. It was shown or highlighted that a choice of unit of analysis assists in the interpretation of results; so as to avoid some fallacies. Spill-over effects can biase the outcome of the trial. Implementation of mixed models or causal models in Stata accounts for clustering effect.

Chapter 4

Data Analysis and Interpretation

*****

Randomisation by group accompanied by an analysis appropriate to randomisation by individual is an exercise in self-deception, however, and should be discouraged.

Cornfield (1978)

Chapter Preview

Up to now we have seen the theoretical background to analysis of clustered data, and the limitations of ordinary regression models that assume observations to be independent. Also, we outlined the limitations of descriptive statistics in the analysis not only of clustered data but of group randomised trials; for descriptive statistics cannot adjust for other covariates. This chapter presents the results of the application of causal models (hierarchical models) using Stata. We present summary statistics, results from unadjusted models, and finally the results from adjusted models. The data used in the analysis is called Grade 12 *data, supplied by Umalusi, a Council for quality assurance in General and Further Education and Training (GFET).*

4.1 Introduction

This study follow the quantitative approach; as such the results in this study are based on statistics. In another words, only statistical results are interpreted. And where inference is performed - the interpretation of results is performed at 95% confidence limit or the interpretation is performed at 0.05 error rate. Therefore, the results that emanated from the application of causal models to the Grade 12 data are declared significant if the probability value (p-value) is less than 0.05. The analysis was performed in two parts. First, descriptive analysis is performed. Second, inferential statistics is performed through the application of causal models.

Grade 12 data are the observations about the learners who sat for Grade 12 examinations in 2008, 2009 and 2010. Appendix A3 on Page 142 presents the definitions of all variables contained in the dataset (Data Dictionary). This dataset is a standardised set because “adjustment methods” were applied on the data. Thus, the dataset contains about 95% of the total learners who wrote the examination[1].

Umalusi Council sets and monitors standards for GFET in South Africa in accordance with the National Qualifications Framework Act No 67 of 2008 and the General and Further Education and Training Quality Assurance Act No 58 of 2001.

The Council is tasked with the development and management of a sub-framework of qualifications for GFET and for the attendant quality assurance. Among others, Umalusi is responsible for the certification of:

a) Senior Certificate (SC) - continues as a revised qualification for adults.

b) National Senior Certificate (NSC) - replaced the SC in 2008.

4.2 Descriptive Analysis

This section presents the results from the descriptive analysis. First, it provides the distribution of learners according to gender and academic year in terms of frequencies and proportions between the two provinces. Secondly, the distribution of learners in the two provinces is presented in terms of the binary outcome (pass or not pass) per academic year. Finally, the average performances (generated from the continuous outcome *marks*), between the two provinces are presented for academic years 2008 through 2010.

As depicted by Table 4.1, the proportion of female learners in Gauteng Province ranged between 54.48% and 54.99%, while in Western Cape Province it ranged between 56.78% and 57.16%, in 2008 through 2010 academic years. During the period under study, proportions of female learners in Western Cape Province are higher than those in Gauteng Province, hence for male learners, the opposite is the case.

[1] The sample size depends entirely on ”date-stamp”. That is, the sample keeps changing depending on when the data were downloaded; since some learners opt to combine their results after having re-written the examinations.

Table 4.1: Distributions of learners by gender and provinces: 2008, 2009 and 2010 academic years.

	Gender	Gauteng Province		Western C. Province	
		Frequency	Percent	Frequency	Percent
2008	Male	44940	45.44	19268	42.84
	Female	53954	54.56	25709	57.16
	Total	**98894**	**100.00**	**44977**	**100.00**
2009	Male	45315	45.52	19930	43.11
	Female	54234	54.48	26304	56.89
	Total	**99549**	**100.00**	**46234**	**100.00**
2010	Male	42475	45.01	20232	43.22
	Female	51893	54.99	26584	56.78
	Total	**94368**	**100.00**	**46816**	**100.00**

Table 4.2 shows that at least 70.42% of learners in Gauteng Province and at least 73.96% of learners in Western Cape Province passed Grade 12 during the years 2008 to 2010. Gauteng Province recorded proportions of learners who passed Grade 12 as 74.33%, 70.42% and 76.90% in 2008, 2009 and 2010, respectively. Similarly Western Cape Province recorded proportions of learners who passed Grade 12 as 77.08%, 73.96% and 74.43% in 2008, 2009 and 2010, respectively.

Table 4.2: Pass rates by provinces: 2008, 2009 and 2010 academic years.

	Result	Gauteng Province		Western C. Province	
		Frequency	Percent	Frequency	Percent
2008	Fail	25381	25.67	10309	22.92
	Pass	73512	74.33	34668	77.08
2009	Fail	29443	29.58	12040	26.04
	Pass	70106	70.42	34194	73.96
2010	Fail	21797	23.10	11974	25.57
	Pass	72580	76.90	34856	74.43

The proportion of learners who passed Grade 12 was marginally higher for Gauteng Province than for Western Cape Province in 2010 (76.90% $[n = 72580]$ vs. 74.43% $[n = 34856]$). Otherwise, marginal proportions of learners who passed Grade 12 favoured Western Cape Province in 2008 (77.08% $[n = 34668]$ vs.74.33% $[n = 73512]$) and in 2009 (73.96% $[n = 34194]$ vs.70.42% $[n = 70106]$). Detailed information is given by Table A.2 on Page 145.

Table 4.3 shows that the average performances for Gauteng Province and Western Cape Province in 2008, 2009 and 2010 are 49.52%, 49.23% and 51.24%, and 50.37%, 50.24 and 50.96%, respectively. Therefore, the marginal average performances favoured the Western Cape Province in 2008 and 2009, while Gauteng Province had marginally higher average score than Western Cape Province in 2010.

Table 4.3 indicates that within standard deviations for marks, as expressed in percentages, over the years are different from zero. This is because, within each learner, the values of this variable (do) vary, i.e. for each of the records the learner has, the values of this variable are different. Also, the between subjects standard deviation is different from 0. This is because all learners have different set of values on marks.

Table 4.3: Average performances per provinces:
2008, 2009 and 2010 academic years

		2008		2009		2010	
		Mean	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Gauteng	overall	49.5203	17.91434	49.22978	18.15893	51.24318	17.70470
	between		12.54402		12.16406		12.25954
	within		12.75883		13.48440		12.82948
Western Cape	overall	50.36789	17.58325	50.23749	17.72550	50.96452	17.88011
	between		13.34398		12.77441		12.86035
	within		11.51861		12.30528		12.42091

In 2008 and 2010 the within and between standard deviations are almost the same for Gauteng Province. This implies that the variation in marks across learners is nearly equal to that observed within a learner over different study subjects (learning areas). That is, if one was to draw two learners randomly from the data, the difference in marks is expected to be nearly equal to the difference for the same learner in two randomly selected any other learning subjects. Similarly are within and between standard deviations for Gauteng Province in 2010, and the within and between standard deviations for Western Cape Province in 2009, and in 2010.

The proportions of female learners were marginaly higher in Western Cape Province than in Gauteng Province in 2008 ($57.16\% - 54.56\% = 2.6\%$), 2009 ($56.89\% - 54.48\% = 2.41\%$) and in 2010 ($56.78\% - 54.99\% = 1.79\%$).

The marginal differences of proportions of learners who passed Grade 12 favoured Western Cape Province in 2008 ($77.08\% - 74.44\% = 2.75\%$) and 2009 ($73.96\% -$

$70.42\% = 3.54\%$). The difference of 2.47% favoured Gauteng Province in 2010.

The marginal average differerences of 0.84759 and 1.00771 favoured the Western Cape Province in 2008 and 2009 academic years, respectively. However, the marginal average difference of 0.27866 favoured Gauteng Province than Western Cape Province in 2010.

4.3 Inferential Statistics

This section presents the results of the inferential statistics. The interpretations were performed at 95% confidence limit (2-sided). The unadjusted and adjusted models were fitted to compare the two provinces. Where Pearson's chi-square test was used - the analysis was performed at $\alpha = 0.05$ error rate.

4.3.1 Analysis of Categorical Variables

The Pearson chi-square statistic tests whether the two categorical variables are independent (Bergsma et al., 2009, Page 17). That is, we use Pearson chi-square test to test for association between two categorical variables (Field, 2005, Page 689), which are *Assessment Outcome* and *Province*. The results are declared significant if $p < 0.05$. The null and alternative hypotheses are:

H_0: Assessment outcome and province are not significantly associated.

H_1: Assessment outcome and provice are significantly associated.

What we mean by association is that the pattern of responses (i.e. the proportion of learners in Gauteng Province to the proportion of learners in the Western Cape Province) in the categories of Assessment outcome is significantly different (Field, 2000; Agresti, 2002; Field, 2005).

Figure 4.1 shows the distribution of proportions for learners by Province and Assessment outcome for the academic year 2008. The proportions favoured the Western Cape Province than Gauteng Province in the categories Bachelor (32.40% vs. 29.46%), Diploma (28.58% vs. 27.53%), and NSC (0.05% vs. 0.04%). Otherwise, the proportions favoured Gauteng Province than Western Cape Province in the categories Fail (25.67% vs. 22.92%) and Higher Certificate HC (17.31% vs. 16.04%).

Figure 4.2 indicates that the proportions favoured the Western Cape Province than the Gauteng Province in the categories Bachelor (31.10% vs. 28.58%) and HC (15.24% vs. 14.04%). The proportions favoured Gauteng Province than Western Cape Province in

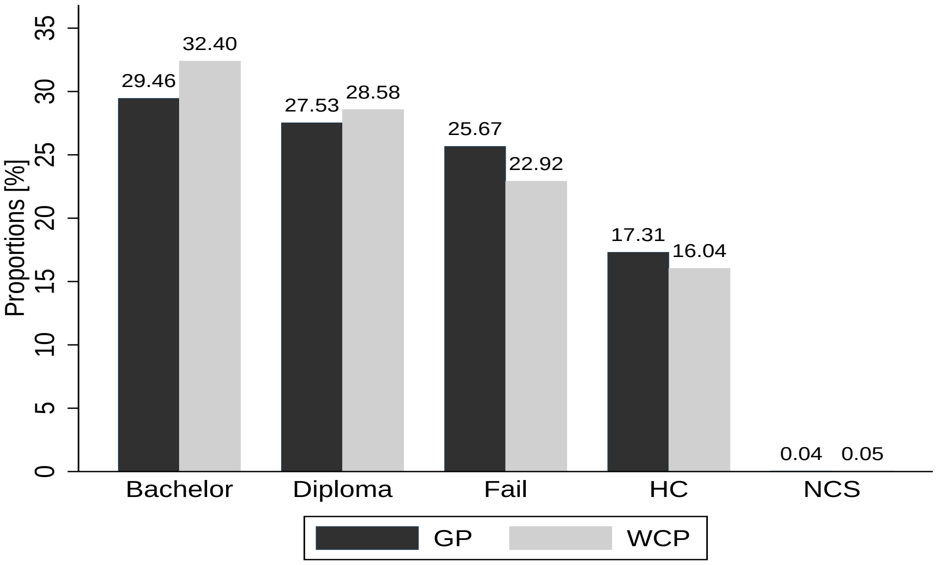


Figure 4.1: Proportions of Assessment Outcome by Province: 2008
GP = Gauteng Province and WCP = Western Cape Province

the categories Diploma (27.78% vs. 27.57%) and Fail (29.58% vs. 26.04%). Otherwise, the proportions were at par at 0.06% under the category of NSC in 2009.

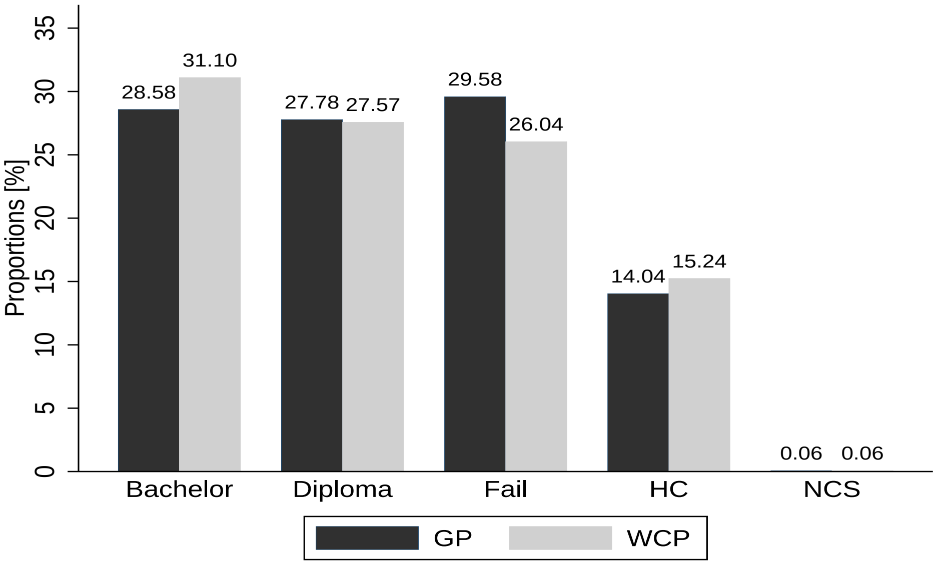


Figure 4.2: Proportions of Assessment Outcome by Province: 2009

Figure 4.3 shows that the proportions favoured the Western Cape Province than Gauteng Province in the categories Fail (25.57% vs. 23.10%), HC (14.85% vs. 13.02%), and NSC (0.06% vs. 0.05%). Otherwise, the proportions favoured Gauteng Province than Western Cape Province in the categories Bachelor (33.17% vs. 30.19%) and Diploma (30.66% vs. 29.33%).

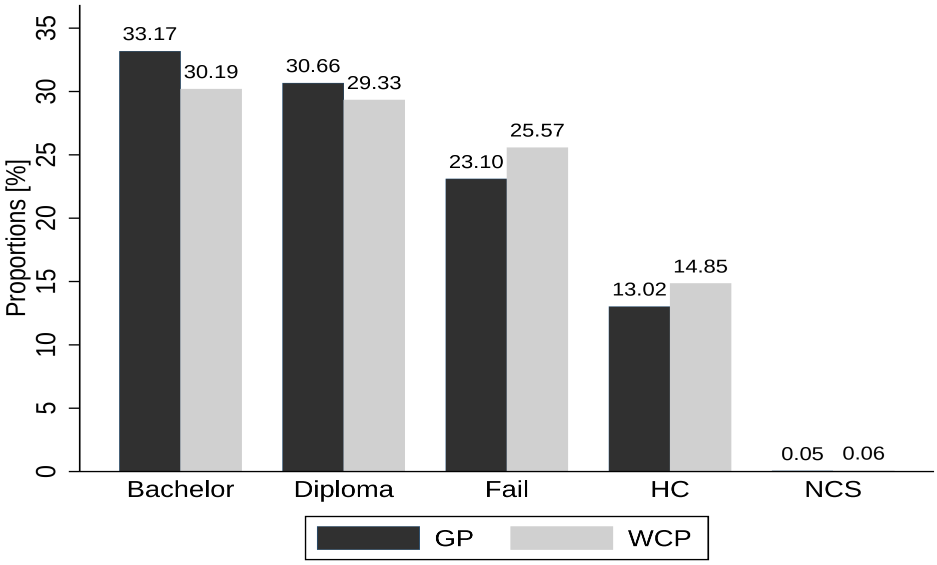


Figure 4.3: Proportions of Assessment Outcome by Province: 2010

The results from Pearson's chi-square tests, as presented by Table 4.4, were found to be highly significant (p-value < 0.001). Therefore, there is enough evidence that the proportions of learners from Gauteng Province to the proportion of learners from Western Cape Province were significantly different in the levels of Assessment Outcome. Therefore, the provinces performed significantly differently in 2008 ($\chi^2_{(4)} = 225.2713, p < 0.001$), 2009 ($\chi^2_{(4)} = 240.3896, p < 0.001$) and in 2010 ($\chi^2_{(4)} = 261.8518, p < 0.001$).

The null hypotheses of no significant association between Assessment outcome and Province for 2008 through 2010 are not accepted.

4.3.2 Application of Causal Models or Hierarchical Models

A classical method for estimating parameters of statistical models is maximum likelihood (Baum, 2006). In particular, we estimate the parameter of the variance-component model using Stata's *xtreg* with *mle* option. The fixed-effect model[2] will be applied to Grade 12 data in order to predict the final marks. The choice of utilising fixed-effects models is due to the fact that there is a lot of within-subject variability or conversely, fixed effect models do not work well when subjects change little across time or occasions (Baum, 2006).

[2] A fixed effects model is a statistical model that represents the observed quantities in terms of explanatory variables that are treated as if the quantities were nonrandom.

Table 4.4: Cross Classification of Assessment Outcome by Province
Note: *Proportions [%] in Parentheses*

	Province	Assessment Outcome					Total
		Bachelor	Diploma	Fail	HC	NSC	
2008	Gauteng	29134	27223	25381	17118	37	**98893**
		(29.46)	(27.53)	(25.67)	(17.31)	(0.04)	**(100.00)**
	Western Cape	14572	12856	10309	7216	24	**44977**
		(32.40)	(28.58)	(22.92)	(16.04)	(0.05)	**(100.00)**
	Total	**43706**	**40,079**	**35690**	**24,334**	**61**	**143870**
		(30.38)	**(27.86)**	**(24.81)**	**(16.91)**	**(0.04)**	**(100.00)**
		(Pearson $\chi^2_{(4)} = 225.2713$; $Pr = 0.000$)					
2009	Gauteng	28417	27654	29443	13980	55	**99549**
		(28.58)	(27.78)	(29.58)	(14.04)	(0.06)	**(100.00)**
	Western Cape	14377	12745	12040	7044	28	**46234**
		(31.10)	(27.57)	(26.04)	(15.24)	(0.06)	**(100.00)**
	Total	**42794**	**40339**	**41483**	**21024**	**83**	**145783**
		(29.35)	**(27.71)**	**(28.46)**	**(14.42)**	**(0.06)**	**(100.00)**
		(Pearson $\chi^2_{(4)} = 240.3896$; $Pr = 0.000$)					
2010	Gauteng	31301	28938	21797	12290	51	**94377**
		(33.17)	(30.66)	(23.10)	(13.02)	(0.05)	**(100.00)**
	Western Cape	14138	13734	11974	6955	29	**46830**
		(30.19)	(29.33)	(25.57)	(14.85)	(0.06)	**(100.00)**
	Total	**45439**	**42672**	**33771**	**19245**	**80**	**144207**
		(32.18)	**(30.22)**	**(23.92)**	**(13.63)**	**(0.06)**	**(100.00)**
		(Pearson $\chi^2_{(4)} = 261.8518$; $Pr = 0.000$)					

4.3.2.1 Continuous Outcome: Marks (%)

The section tests the null hypothesis that the average marks (%) between the two provinces are not different. First, the unadjusted models were considered. Secondly, the adjusted models were fitted to check the effect of gender, quintile, then gender and quintile.

The null hypothesis (H_0) and alternative hypothesis (H_1) are given as:

$$H_0 : \hat{\mu}_{GP} = \hat{\mu}_{WCP}$$

$$H_1 : \hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$$

where $\hat{\mu}_{GP}$ and $\hat{\mu}_{WCP}$ are estimated averages for Gauteng Province and Western Cape Province, respectively.

A. **Unadjusted Model**

The hypotheses that the average marks in the two provinces are not different are tested agains the alternative hypotheses that the average marks are different.

The results from unadjusted models for academic years 2008, 2009 and 2010 are given by Table 4.5, Table 4.6 and Table 4.7, respectively. The estimates from unadjusted models are crude estimates.

The crude estimate, as depicted by Table 4.5, indicates that Western Cape Province performed significantly better than Gauteng Province in 2008. In particular, for every percentage increase in *marks obtained*, learners in Western Cape Province were expected to score 0.782% more than the learners in Gauteng Province.

Table 4.5: Comparison of Provinces in 2008

Variables	marks	sigma_u	sigma_e
Western Cape Province	0.782^{***}		
	(0.0727)		
Constant	49.41^{***}	11.71^{***}	13.38^{***}
	(0.0406)	(0.0260)	(0.0102)
Observations	$1,003,954$	$1,003,954$	$1,003,954$
Number of candidate	$143,870$	$143,870$	$143,870$

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Gauteng Province is a Reference Province

Therefore, the hypothesis of no average difference between the study provinces was not accepted in favour of alternative hypothesis that the average marks between the two provinces were different. Thus, Western Cape Province performed significantly better than Gauteng Province in 2008 academic year.

The unadjusted estimate as shown in Table 4.6 indicates that Western Cape Province performed significantly better than Gauteng Province in 2009. For every percentage increase in *marks obtained*, learners from Western Cape Province are expected to score 0.957% more than the learners from Gauteng Province.

The hypothesis of no average difference is not accepted in favour of alternative hypothesis that the average marks between the two provinces are different. Thus, Western Cape Province performed significantly better than Gauteng Province in 2009 academic year.

Table 4.7 presents the results of the unadjusted regression model. The crude estimate indicates that the performance of the Western Cape Province was significantly different from Gauteng Province. In particular, Western Cape Province

Table 4.6: Comparison of Provinces in 2009

Variables	marks	sigma_u	sigma_e
Western Cape Province	0.957^{***}		
	(0.0694)		
Constant	49.13^{***}	11.08^{***}	14.17^{***}
	(0.0391)	(0.0254)	(0.0107)
Observations	$1,018,728$	$1,018,728$	$1,018,728$
Number of candidate	$145,783$	$145,783$	$145,783$

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Gauteng Province is a Reference Province

was expected to score a significant final mark of 0.302 less than Gauteng Province, with every 1% increase in marks obtained.

Table 4.7: Comparison of Provinces in 2010

Variables	marks	sigma_u	sigma_e
Western Cape Province	-0.302^{***}		
	(0.0701)		
Constant	51.10^{***}	11.23^{***}	13.72^{***}
	(0.0404)	(0.0258)	$0.0105)$
Observations	$985,956$	$985,956$	$985,956$
Number of candidate	$141,207$	$141,207$	$141,207$

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Gauteng Province is a Reference Province

Therefore, the hypothesis of no average difference is not accepted in favour of alternative hypothesis that the average marks between the two provinces are different. Thus, learners from the Western Cape Province performed significantly better than learners from the Gauteng Province in 2010 academic year.

The (null) hypotheses that the average performances between the two provinces were not significantly different in 2008 through 2010 academic years are not accepted. The average differences favoured the Western Cape Province. In other words, learners from Western Cape Province were expected to significantly score higher marks than learners from Gauteng Province in 2008, 2009 and 2010 academic years.

B. **Adjusted Models**

This section deals with the comparisons of the two provinces after controlling for other covariates. That is, we want to check or determine how provinces differed after adjusting for quintile, gender, and quintile and gender.

a) **Adjusted Models for** 2008

The comparisons between the two provinces is determined after adjusting for quintile, gender, and quintile and gender. In another words, we compare the average marks (%) between the two provinces after controlling for the said covariates.

i) **Adjusting for Quintile**

The hypothesis of no average differences between the two provinces after adjusting for quintile is tested. To be precise, the null hypothesis and alternative hypothesis are:

(H_0): $\hat{\mu}_{GP} = \hat{\mu}_{WCP}$ given that quintile is constant.

(H_1): $\hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$ given that quintile is constant.

The result indicates that learners from Western Cape Province were expected to significantly score higher marks than learners from Gauteng Province. Strictly speaking, Western Cape Province is expected to significantly score 2.091 final marks more than Gauteng Province given that quintile was constant. Table 4.8 summarises the output of regression model, and makes this explanation more explicit.

Therefore, there is a significant difference in average marks between Gauteng Province and Western Cape Province when the Quintile is kept constant. In particular, the average difference favoured Western Cape Province after adjusting for quintile.

Therefore, hypothesis of no average difference between the two provinces was not accepted. Thus, we accepted (H_1), the hypothesis that Gauteng Province and Western Cape Province performed significantly differently after adjusting for Quintile."

ii) **Adjusting for Gender**

The null hypothesis is stated as: There is no significant average difference in marks (%) between the two provinces after adjustibg for gender. To be precise, we state the null hypothesis (H_0) and alternative hypothesis (H_1) as:

(H_0): $\hat{\mu}_{GP} = \hat{\mu}_{WCP}$ given that Gender is constant.

Table 4.8: Comparison of Provinces : 2008
Adjusting for Quintile

variable	marks	sigma_u	sigma_e
Western Cape Province	2.091^{***}		
	(0.0645)		
quintile_2	-0.154		
	(0.127)		
quintile_3	1.864^{***}		
	(0.116)		
quintile_4	4.093^{***}		
	(0.114)		
quintile_5	15.75^{***}		
	(0.108)		
Constant	42.06^{***}	9.617^{***}	13.38^{***}
	(0.102)	(0.0231)	(0.0102)
Observations	$994,353$	$994,353$	$994,353$
Number of candidate	$142,483$	$142,483$	$142,483$

Standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1

Gauteng Provive is a Reference Province

Quintile_1 is a Reference quintile

(H_1): $\hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$ given that Gender is constant.

The result from adjusted model, as presented by Table 4.9, indicates that learners from Western Cape Province have significantly performed better than learners from Gauteng Province after controlling for gender. That is, learners from Western Cape Province were expected to significantly score 0.746 more final marks than learners from Gauteng Province when gender was kept constant.

The hypothesis that the difference between the two provinces was not significant was rejected. Therefore, the two provinces performed significantly different. In particular, the Western Cape Province performed significantly better than Gauteng Province after adjusting for gender.

iii) **Adjusting for both Gender and Quintile**

The hypothesis that Western Cape Province and Gauteng Province did not performe significantly different after adjusting for both gender and quintile was tested. The null and alternative hypotheses are:

Table 4.9: Comparison of Provinces : 2008
Adjusting for Gender

Variables	Marks	sigma_u	sigma_e
Western Cape Province	0.746***		
	(0.0726)		
Gender	1.391***		
	(0.0677)		
Constant	48.65***	11.69***	13.38***
	(0.0548)	(0.0260)	(0.0102)
Observations	$1,003,954$	$1,003,954$	$1,003,954$
Number of candidate	$143,870$	$143,870$	$143,870$

Standard errors in parentheses
*** p < 0.01, ** p < 0.05, * p < 0.1
Gauteng Provive is a Reference Province

(H_0): $\hat{\mu}_{GP} = \hat{\mu}_{WCP}$ given that both learner gender and school quintile are constant.

(H_1): $\hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$ given that both learner gender and school quintile are constant.

The results are presented by Table 4.10. It can be read that learners from Western Cape Province significantly performed better than learners from Gauteng Province. After controlling for gender and quintile; we expected learners from Western Cape Province to score 2.051 marks more than learners from Gauteng Province ($p < 0.01$).

The hypothesis of insignificant differences between the two provinces after adjusting for both quintile and gender was rejected, and the Western Cape Province performed significantly better than Gauteng Province.

Summary

The Western Cape Province performed significantly better than Gauteng Province in 2008. Both unadjusted and adjusted models (here, adjusting for gender, quintile, and gender and quintile) indicated that Western Cape Province performed significantly better than Gauteng Province. The adjusted models indicate that Western Cape Province was expected to score at least 0.7 points more than Gauteng Province with every 1% increase in final mark.

We note that if schools had same resources (belonged to the same quintile)

Table 4.10: Comparison of Provinces : 2008
Adjusting for Quintile and Gender

Variables	marks	sigma_u	sigma_e
Western Cape Province	2.051***		
	(0.0643)		
Female	1.700***		
	(0.0579)		
quintile_2	-0.157		
	(0.127)		
quintile_3	1.889***		
	(0.116)		
quintile_4	4.091***		
	(0.114)		
quintile_5	15.80***		
	(0.108)		
Constant	41.11***	9.580***	13.38***
	(0.107)	(0.0231)	(0.0102)
Observations	994,353	994,353	994,353
Number of candidate	142,483	142,483	142,483

Standard errors in parentheses
*** p < 0.01, ** p < 0.05, * p < 0.1
Gauteng Province is a Reference province
Quintile_1 is a Reference quintile

then learners from the Western Cape Province would significantly perform better than learners from Gauteng Province. Also, if learners from Gauteng Province and Western Cape Province belonged to the same gender-category (same sex) then one would expect learners from Western Cape Province to perform significantly better than learners from Gauteng Province. Similarly, if all schools were equally resourced and learners belonged to the same gender-category then one would expect learners from Western Cape Province to perform significantly better than learners from Gauteng Province.

Therefore, there is a sufficient evidence to support the fact that Western Cape Province performed significantly better than Gauteng Province in 2008 academic year.

b) Adjusted models for 2009

This section presents the results of the adjusted models as fitted to the 2009 data. The average marks (%) between the provinces are compared after adjusting for Quintile, Gender, Gender and Quintile.

i) **Adjusting for Quintile**

The hypothesis of no average differences between the two provinces after adjusting for quintile is tested. Here, the null hypothesis (H_0) and alternative hypothesis (H_1) are:

(H_0): $\hat{\mu}_{GP} = \hat{\mu}_{WCP}$ given that quintile is constant.

(H_1): $\hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$ given that quintile is constant.

Table 4.11 presents the results of the regression model adjusting for Quintile. The learners from Western Cape Province had significantly performed better than learners from Gauteng Province ($p < 0.01$). With every 1% increase in marks obtained learners from Western Cape Province were expected to score 2.268 marks more than the learners from Gauteng Province.

Table 4.11: Comparison of Provinces : 2009
Adjusting for Quintile

Variables	marks	sigma_u	sigma_e
Western Cape Province	2.268^{***}		
	(0.0631)		
quintile_2	0.814^{***}		
	(0.126)		
quintile_3	2.048^{***}		
	(0.113)		
quintile_4	3.869^{***}		
	(0.112)		
quintile_5	14.70^{***}		
	(0.105)		
Constant	41.98^{***}	9.276^{***}	14.17^{***}
	(0.0994)	(0.0232)	(0.0108)
Observations	$1,005,222$	$1,005,222$	$1,005,222$
Number of candidate	$143,828$	$143,828$	$143,828$

Standard errors in parentheses
*** p < 0.01, ** p < 0.05, * p < 0.1
Gauteng Province is a Baseline province
Quintile_1 is a Baseline quintile

The hypothesis that the average marks (%) between the provinces after adjusting for quintile are not different isrejected. Thus, there is a significant difference in marks between Gauteng Province and Western Cape Province after controlling for quintile; and the difference favoured the Western Cape Province.

ii) **Adjusting for Gender**

The hypothesis of no average differences between the two provinces after adjusting for gender is tested. The null hypothesis (H_0) and alternative hypothesis (H_1) are:

(H_0): $\hat{\mu}_{GP} = \hat{\mu}_{WCP}$ given that gender is constant.
(H_1): $\hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$ given that gender is constant.

Table 4.12 indicates that learners from Western Cape Province had performed significantly better than learners from Gauteng Province ($p < 0.01$). For every 1% increase in marks obtained one would expect learners from Western Cape Province to score 0.921 marks more than the learners from Gauteng Province.

Table 4.12: Comparison of Provinces : 2009
Adjusting for Gender

Variables	marks	sigma_u	sigma_e
Western Cape Province	0.921***		
	(0.0693)		
Gender	1.488***		
	(0.0649)		
Constant	48.32***	11.06***	14.17***
	(0.0526)	(0.0254)	(0.0107)
Observations	$1,018,728$	$1,018,728$	$1,018,728$
Number of candidate	$145,783$	$145,783$	$145,783$

Standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1
Gauteng Province is a Baseline province

Therefore, the hypothesis (H_0) is rejected, and we conclude that there is enough evidence that learners from Western Cape Province performed significantly better than learners from Gauteng Province given that gender is constant.

iii) **Adjusting for Gender and Quintile**

The hypothesis that Gauteng Province and Western Cape Province did perform differently after controlling for gender and quintile is tested against the alternative hypothesis that the average performances between the two provinces are different.

In another words, the null hypothesis (H_0) and the alternative hypothesis (H_1) are:

(H_0): $\hat{\mu}_{GP} = \hat{\mu}_{WCP}$ given that learner gender and school quintile are constant.

(H_1): $\hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$ given that learner gender and school quintile are constant.

Learners from Western Cape Province scored significantly higher marks than learners from Gauteng Province ($p < 0.01$). In particular, as depicted by Table 4.13, for every 1% increase in marks obtained one would expect learners from Western Cape Province to significantly score 2.231 marks more than the learners from Gauteng Province.

There is sufficient evidence that learners from Western Cape Province performed significantly better than learners from Gauteng Province. Therefore, the null hypothesis was rejected.

Summary

As in 2008, the Western Cape Province performed significantly better than Gauteng Province in 2009. Both the crude and adjusted models indicate that Western Cape Province performed significantly better than Gauteng Province. The adjusted models indicate that Western Cape Province was expected to score at least 2% points significantly more than Gauteng Province with every 1% increase in final mark.

We note that if all schools were resourced equally then learners from the Western Cape Province would perform significantly better than learners from Gauteng Province. Also, if learners from Gauteng Province and Western Cape Province belonged to the same gender-category (same sex) then one would expect learners from Western Cape Province to perform significantly better than learners from Gauteng Province.

Similarly, if all schools were equally resourced and learners belonged to the same gender-category then one would expect learners from Western Cape Province to perform significantly better than learners from Gauteng Province.

Therefore, there is a sufficient evidence to support the fact that Western

Table 4.13: Comparison of Provinces : 2009
Adjusting for Gender and Quintile

Variables	marks	sigma_u	sigma_e
Western Cape Province	2.231***		
	(0.0629)		
Gender	1.765***		
	(0.0568)		
quintile_2	0.805***		
	(0.126)		
quintile_3	2.071***		
	(0.113)		
quintile_4	3.869***		
	(0.112)		
quintile_5	14.75***		
	(0.104)		
Constant	40.99***	9.235***	14.17***
	(0.104)	(0.0232)	(0.0108)
Observations	1,005,222	1,005,222	1,005,222
Number of candidate	143,828	143,828	143,828

Standard errors in parentheses
*** p < 0.01, ** p < 0.05, * p < 0.1
Gauteng Provice is a reference province
Quintile_1 is a reference quintile

Cape Province performed significantly better than Gauteng Province in 2009 academic year.

c) **Adjusted models for 2010**

This section gives the results of the adjusted models for 2010 academic year. We compare the average marks between the two provinces after controlling for gender, quintile, and gender and quintile.

i) **Adjusting for Quintile**

We test the null hypothesis (H_0) against the alternative hypothesis (H_1) which are stated as:

(H_0): The average marks between the two provinces are not different given that quintile is constant,

(H_1): The average marks between the two provinces are significantly different given that quintile is constant.

Learners from Western Cape Province had significantly higher scores than learners from Gauteng Province ($p < 0.01$) after adjusting for school quintile. In another words, learners from Western Cape Province are expected to significantly score about 1.30 more than the learners from Gauteng Province with every 1% increase in marks obtained ($p < 0.01$). Table 4.14 summarises the output of the fitted model.

Table 4.14: Comparison of Provinces : 2010
Adjusting for Quintile

Variables	**marks**	**sigma_u**	**sigma_e**
Western Cape Province	1.290***		
	(0.0650)		
quintile_2	0.545***		
	(0.129)		
quintile_3	1.633***		
	(0.116)		
quintile_4	3.265***		
	(0.114)		
quintile_5	13.67***		
	(0.106)		
Constant	44.37***	9.656***	13.71***
	(0.100)	(0.0238)	(0.0107)
Observations	$965,114$	$965,114$	$965,114$
Number of candidate	$138,189$	$138,189$	$138,189$

Standard errors in parentheses
*** p < 0.01, ** p < 0.05, * p < 0.1
Gauteng Provice is a reference province
Quintile_1 is a reference quintile

The hypothesis that the average marks between the two provinces are not different after adjusting for school quintile is rejected. Therefore, there is a sufficient evidence that learners from Western Cape Province performed significantly better than learners from Gauteng Province after controlling for school quintile.

ii) **Adjusting for Gender**

The null hypothesis of no average differences between the two provinces after adjusting for gender is tested. The null hypothesis (H_0) and alternative hypothesis (H_1) are:

(H_0): $\hat{\mu}_{GP} = \hat{\mu}_{WCP}$ given that gender is constant.

(H_1): $\hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$ given that gender is constant.

Learners from Western Cape Province had significantly lower scores than learners from Gauteng Province after adjusting for gender. In another words, learners from Western Cape Province were expected to significantly score about 0.33 marks less than the learners from Gauteng Province when gender is kept constant ($p < 0.01$). Table 4.15 presents the summarised results.

Table 4.15: Comparison of Provinces : 2010
Adjusting for Gender

Variables	marks	sigma_u	sigma_e
Western Cape Province	-0.326***		
	(0.0700)		
Female	1.335***		
	(0.0663)		
Constant	50.37***	11.21***	13.72***
	(0.0543)	(0.0257)	(0.0105)
Observations	$985,817$	$985,817$	$985,817$
Number of candidate	$141,184$	$141,184$	$141,184$

Standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1
Gauteng Provice is a reference province
Male is a reference level of gender

The hypothesis that the average marks between the two provinces are not statistically different is rejected. Therefore, we have sufficient evidence to conclude that after adjusting for gender - learners from Gauteng Province performed sgnificantly better that learners from Western Cape Province in 2010.

iii) **Adjusting for Gender and Quintile**

We test the null hypothesis of no average differences between the two provinces after adjusting for gender and school quintile. The null hypothesis (H_0) and alternative hypothesis (H_1) are stated as:

H_0: $\hat{\mu}_{GP} = \hat{\mu}_{WCP}$ after adjusting for gender and school quintile.

H_1: $\hat{\mu}_{GP} \neq \hat{\mu}_{WCP}$ after adjusting for gender and quintile.

The result from adjusted model indicates that learners from Western Cape Province were significantly different from those in Gauteng Province

($p < 0.01$). Adjusting for both Gender and Quintile, (see Table 4.16), the learners from Western Cape Province were expected to score about 1.27 marks more than learners from Gauteng Province.

Table 4.16: Comparison of Provinces : 2010
Adjusting for Quintile and Gender

Variables	marks	sigma_u	sigma_e
Western Cape Province	1.268***		
	(0.0648)		
Female	1.648***		
	(0.0593)		
quintile_2	0.550***		
	(0.129)		
quintile_3	1.669***		
	(0.116)		
quintile_4	3.284***		
	(0.114)		
quintile_5	13.72***		
	(0.106)		
Constant	43.43***	9.621***	13.71***
	(0.106)	(0.0238)	(0.0107)
Observations	964,994	964,994	964,994
Number of candidate	138,169	138,169	138,169

Standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1
Gauteng Provice is a reference province
Male is a reference level of gender
Quintile_1 is a reference quintile

The hypothesis that the average marks between the two provinces were not statistically different is rejected. Therefore, we have sufficient evidence to conclude that after adjusting for gender and quintile then learners from Western Cape Province performed significantly better than learners from Gauteng Province in 2010.

Summary

All adjusted models indicated that learners from Western Cape Province performed significantly better than learners from Gauteng Province with the exception of when the adjustment was made for only gender.

Therefore, if all schools were equally resourced and all learners were of the same gender then learners from Western Cape Province would score significantly higher marks than learners from Gauteng Province. Similarly, if all schools were equally resourced then one would expect learners from Western Cape Province to score significantly higher marks than learners from Gauteng Province. The adjustment for gender only suggests that learners from Gauteng Province would score significantly higher marks than learners from Western Cape Province.

4.3.2.2 Binary outcome: Final (Fail [not promoted] or Pass [promoted])

The two provinces, Western Cape and Gauteng, are compared over the dichotomous outcome, *final*, which indicates whether or not a learner passed Grade 12. The likelihood of observing a pass (*promoted*) or not pass (*not promoted*) between the two provinces is determined. The parameter of interest is odds ratio (OR). The logistic regression models were fitted for 2008, 2009 and 2010 datasets. The models account for *ICC*.

A) **Unadjusted Models**

The section tests the null hypothesis that the average marks (%) between the two provinces are not different. Firstly, the unadjusted models were considered. Secondly, the adjusted models were fitted to check the effect of gender, quintile, and gender and quintile.

The null hypothesis which says that the proportions in the two provinces are not different (H_0) and alternative hypothesis which says that the proportions in the two provinces are significantly different (H_1) are given as:

$$H_0 : \hat{p}_{GP} = \hat{p}_{WCP}$$
$$H_1 : \hat{\mu}_{GP} \neq \hat{p}_{WCP}$$

where $\hat{p}_{GP}$ and $\hat{p}_{WCP}$ are estimated proportions for Gauteng Province and Western Cape Province, respectively.

a) **Unadjusted Model:** 2008

We test the hypothesis that the proportions in the two provinces are not significantly different (H_0) against the hypothesis that the proportions are significantly different (H_1).

The two provinces are significantly different since learners from Western Cape Province were about 1.16 more likely than learners from Gauteng Province to pass Grade 12 in 2008. As depicted by Table 4.17, the parameters of interest are given as $p < 0.01, OR = 1.161126$, and 95%

$CI = (1.131252 - 1.191788)$. Therefore, the odds of passing Grade 12 increased by a factor of 1.16 for learners in the Western Cape Province over that of learners in Gauteng Province.

Table 4.17: Comparison of Provinces : 2008
Crude Estimates

Final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]	
Western Cape						
Province	1.161126	$.0154412$	11.23	0.000	1.131252	1.191788
cons	2.897499	$.0209761$	146.95	0.000	2.856677	2.938904
Gauteng Provice is a reference province						

Therefore, the hypothesis that the proportion of learners who passed Grade 12 to those who failed Grade 12 in the two provinces is the same is rejected. We have sufficient evidence that learners in the Western Cape Province were significantly more likely than learners in Gauteng Province to pass Grade 12 in 2008 academic year.

b) **Unadjusted Model:** 2009

We test the hypothesis that the proportions in the two provinces are not significantly different (H_0) against the hypothesis that the proportions are significantly different (H_1).

Learners from Western Cape Province than learners from Gauteng Province were about 1.19 more likely to pass Grade 12 ($p < 0.01, OR = 1.192776, 95\%$ $CI : 1.163649 - 1.222632$). In another words, the odds of passing Grade 12 increased significantly by a factor of about 1.19 for learners from Western Cape Province over that of learners from Gauteng Province, as can be seen from Table 4.18.

Table 4.18: Comparison of Provinces : 2009
Crude Estimates

Final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]	
Western Cape						
Province	1.192776	$.0150453$	13.98	0.000	1.163649	1.222632
cons	2.381737	$.0164666$	125.52	0.000	2.349681	2.414231
Gauteng Provice is a reference province						

Therefore, the hypothesis that the proportion of learners who passed Grade 12 to the proportion of learners who failed Grade 12 in the two provinces was the same is rejected. We have sufficient evidence that learners in the Western Cape Province were significantly more likely than learners in Gauteng Province to pass Grade 12 in 2009 academic year.

c) **Unadjusted Model:** 2010

The unadjusted model was fitted to test for difference in proportions between the two provinces. In particular, the null hypothesis of no difference in proportions was tested against the alternative hypothesis that the proportions between the two provinces are significantly different.

Learners from Western Cape Province than learners from Gauteng Province were about 0.87 less likely to pass Grade 12 ($p < 0.01, OR = 0.8741196, 95\%$ $CI : 0.8520777 - 0.8967317$). In another words, the odds of passing Grade 12 decreased by a factor of about 0.87 for learners from Western Cape Province over that of learners from Gauteng Province (as can be seen from Table 4.19).

Table 4.19: Comparison of Provinces : 2010
Crude Estimates

Final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]	
Western Cape						
Province	$.8741196$	$.0113903$	-10.32	0.000	$.8520777$	$.8967317$
cons	3.331823	$.0255799$	156.76	0.000	3.282063	3.382338
Gauteng Provice is a reference province						

The hypothesis of no difference in proportions between the two provinces is rejected. That is, the proportion of learners who passed Grade 12 in Western Cape Province was significantly lower than that of those in Gauteng Province. Therefore, learners in Gauteng Province were significantly more likely than learners in Western Cape Province to pass Grade 12 in 2010.

Summary

The crude estimates indicate that learners from Western Cape Province were significantly more likely than learners from Gauteng Province to pass Grade 12 in 2008 and in 2009. However, the odds of passing Grade 12 favoured learners from Gauteng Province than learners from Western Cape Province in 2010.

B) **Adjusted Model**

This section presents the results of adjusted models when the outcome is a binary outcome, *final*. First, the model will adjust for quintile then adjust for gender before adjusting for both gender and quintile. The results are presented in tabular formats, as in the preceding sections.

a) **Adjusted model for** 2008

This section presents the results of the adjusted models as fitted to the 2008 data. The proportions of learners who passed Grade 12 to those who did not pass Grade 12 in the two provinces are compared after adjusting for Quintile, Gender and then both Gender and Quintile.

i) **Adjusting for Quintile**

The hypothesis that the proportions in the two provinces are not different after controlling for school quintile (H_0) is tested against the alternative hypothesis that the proportions in the two provinces are significantly different after adjusting for quintile (H_1). Symbolically,

H_0: $\hat{p}_{GP} = \hat{p}_{WCP}$ after controlling for school quintile.

H_1: $\hat{p}_{GP} \neq \hat{p}_{WCP}$ after controlling for school quintile.

Table 4.20 presents the estimates after adjusting for school quintile. The result indicates that learners from Western Cape Province had significantly more chances than learners from Gauteng Province of passing Grade 12 in 2008 ($p < 0.001$).

To be precise, learners from Western Cape Province than learners from Gauteng Province were about 1.509 more likely to pass Grade 12 in 2008 ($OR = 1.499607$, 95% $CI : 1.457 - 1.544$). Thus the odds of passing Grade 12 increased by a factor of about 1.50 for learners in the Western Cape Province over that of learners in Gauteng Province.

Table 4.20: Comparison of Provinces : 2008
Adjusting for Quintile

Final	Odds Ratio	Std. Err.	z	$P > \lvert z\rvert$	[95% Conf.	Interval]
Western Cape						
Province	1.499607	0.0222416	27.32	0.000	1.456642	1.54384
quintile_2	1.039562	0.0246533	1.64	0.102	0.992349	1.089022
quintile_3	1.506394	0.0333945	18.48	0.000	1.442344	1.573289
quintile_4	2.086933	0.0463103	33.15	0.000	1.998112	2.179703
quintile_5	9.23663	0.2274573	90.28	0.000	8.801409	9.693372
cons	1.144073	0.0224871	6.85	0.000	1.100837	1.189007

Gauteng Provice is a reference province
Quintile_1 is a reference quintile

There is sufficient evidence to support the alternative hypothesis. In another words, we reject the null hypothesis. Therefore, the two provinces performed significantly differently. Hence, learners from Western Cape Province than learners from Gauteng Province had significantly better chances of passing Grade 12.

ii) **Adjusting for Gender**

The hypothesis that the proportions in the two provinces are not different after controlling for gender (H_0) is tested against the alternative hypothesis that the proportions in the two provinces are significantly different after adjusting for gender (H_1). These hypotheses may be stated as:

H_0: $\hat{p}_{GP} = \hat{p}_{WCP}$ after controlling for gender.

H_1: $\hat{p}_{GP} \neq \hat{p}_{WCP}$ after controlling for gender.

Learners from Western Cape Province than learners from Gauteng Province were 1.16 more likely to pass ($p < 0.01, OR = 1.161375, 95\%CI : 1.131487 - 1.192053$). The odds of passing Grade 12 increased by a factor of about 1.16 for learners from Western Cape Province over that of learners from Gauteng Province. Table 4.21 makes this explanation more explicit.

Table 4.21: Comparison of Provinces : 2008
Adjusting for Gender

final	Odds Ratio	Std. Err.	z	$P > \lvert z \rvert$	[95% Conf.	Interval]
Western Cape Province	1.161375	0.015449	11.25	0.000	1.131487	1.192053
Female	0.991784	0.012109	-0.68	0.499	0.9683313	1.015804
constant	2.910582	0.028658	108.51	0.000	2.854953	2.967295

Gauteng Provice is a reference province
Male is a reference gender

The hypothesis of no difference is rejected. Therefore, there is a sufficient evidence that learners from Western Cape Province than learners from Gauteng Province had better chances of passing Grade 12 after adjusting for gender.

iii) **Adjusting for Gender and Quintile**

The hypothesis that the proportions in the two provinces are not different after controlling for gender and school quintile (H_0) is tested against the alternative hypothesis that the proportions in the two provinces are significantly different after adjusting for gender and school quintile (H_1). These hypotheses may be stated as:

H_0: $\hat{p}_{GP} = \hat{p}_{WCP}$ after controlling for gender.

H_1: $\hat{p}_{GP} \neq \hat{p}_{WCP}$ after controlling for gender.

Learners from Western Cape Province than learners from Gauteng Province were 1.50 more likely to pass Grade 12 ($p < 0.01$, $OR = 1.498202$, $95\%\ CI : 1.455254 - 1.542419$). After adjusting for both gender and quintile, as can be seen from Table 4.22, the odds of passing Grade 12 increased by a factor of about 1.50 for learners from Western Cape Province over that of learners from Gauteng Province.

Table 4.22: Comparison of Provinces : 2008
Adjusting for Quintile and Gender

final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
Western Cape	1.498202	.0222332	27.24	0.000	1.455254	1.542419
gender_2	1.023837	.0132263	1.82	0.068	.9982391	1.050091
quintile_2	1.039447	.0246515	1.63	0.103	.9922362	1.088903
quintile_3	1.506705	.0334028	18.49	0.000	1.442639	1.573617
quintile_4	2.086581	.0463038	33.14	0.000	1.997772	2.179337
quintile_5	9.240944	.2275759	90.29	0.000	8.805497	9.697925
constant	1.129395	.0235943	5.82	0.000	1.084085	1.176599

Gauteng Provice is a reference province
Male is a reference gender
Quintile_1 is a reference quintile

There is sufficient evidence that after controlling for gender and school quintile then the proportion of learners who passed to the proportion of learners who did not pass in the two provinces is significantly different. In particular, learners from Western Cape Province than learners from Gauteng Province had better chances to pass Grade 12 when school quintile and learner gender are kept constant. Therefore, the null hypothesis was rejected.

Summary

Both crude estimates and adjusted estimates indicated that learners in the Western Cape Province than learners from Gauteng Province were more likely to pass Grade 12 in 2008.

b) **Adjusted models for 2009**

This section presents the adjusted models as fitted to the 2009 dataset. As with adjusted models for 2008, Quintile, Gender, and then both Quintile and Gender will be adjusted when determining the estimates (in particular, the odds ratios).

i) **Adjusting for Quintile**

The hypothesis that the proportion of learners who passed to those who did not pass in the two provinces are not different is tested against the hypothesis that the proportion of learners who passed to those who did not pass in the two provinces are significantly different after adjusting for school quintile. The null (H_0) and alternative (H_1) hypotheses may be stated as:

H_0: $\hat{p}_{GP} = \hat{p}_{WCP}$ after controlling for school quintile.

H_1: $\hat{p}_{GP} \neq \hat{p}_{WCP}$ after controlling for school quintile.

The learners from Western Cape Province than learners from Gauteng Province were about 1.51 more likely to pass Grade 12 ($p < 0.01, OR = 1.508262, 95\%CI : 1.467188 - 1.550486$). As indicated by Table 4.23, the odds of passing Grade 12 increased by a factor of about 1.51 for learners from Western Cape Province over that of learners from Gauteng Province after keeping quintile constant.

Table 4.23: Comparison of Provinces : 2009
Adjusting for Quintile

final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]	
Western Cape						
Province	1.508262	$.0212471$	29.17	0.000	1.467188	1.550486
quintile_2	1.207525	$.0286442$	7.95	0.000	1.152669	1.264993
quintile_3	1.460334	$.0313882$	17.62	0.000	1.400092	1.523168
quintile_4	1.86023	$.0399798$	28.88	0.000	1.783499	1.9402634
quintile_5	7.748586	$.1760387$	90.12	0.000	7.411126	8.101413
cons	$.9581361$	$.0183191$	-2.24	0.025	$.9228957$	$.9947222$
Quintile_1 is a reference quintile						
Gauteng Province is a reference province						

The hypothesis that the proportions in the two provinces are not significantly different (H_0) is rejected. Therefore, we accept the alternative hypothesis that the proportions in the two provinces are significantly different. We have sufficient evidence to say after controlling for school quintile in 2009 - learners in the Western Cape Province than learners in Gauteng Province were significantly more likely to pass Grade 12.

ii) **Adjusting for Gender**

The hypothesis that the proportion of learners who passed to those who did not pass in the two provinces are not different is tested against the hypothesis that the proportion of learners who passed to those who did not pass in the two provinces are significantly different after adjusting for gender. The null (H_0) and alternative (H_1) hypotheses may be stated as:

H_0: $\hat{p}_{GP} = \hat{p}_{WCP}$ after adjusting for gender.

H_1: $\hat{p}_{GP} \neq \hat{p}_{WCP}$ after adjusting for gender.

The odds of passing Grade 12 in 2009 significantly favoured the learners from Western Cape Province than learners from Gauteng Province.

Table 4.24 indicates that learners from Western Cape Province than learners from Gauteng Province were about 1.19 more likely to pass Grade 12 ($p < 0.01$, $OR = 1.193263$, 95% CI: $1.164117 - 1.223139$). However, gender is not a significant predictor of whether or not a learner will pass ($p = 0.148$).

Table 4.24: Comparison of Provinces : 2009
Adjusting for Gender

final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
Western Cape Province	1.193263	.0150553	14.00	0.000	1.164117	1.223139
Gender	.9833113	.0114418	-1.45	0.148	.9611397	1.005994
cons	2.40371	.0225791	93.36	0.000	2.35986	2.448374
Gauteng Province is a reference province						

We reject the null hypothesis that the proportions in the two provinces are the same. Therefore, we have sufficient evidence that learners from Western Cape Province than learners from Gauteng Province had significantly better chances to pass Grade 12 after adjusting for gender in 2009 academic year.

iii) **Adjusting for Quintile and Gender**

The hypothesis that the proportion of learners who passed to those who did not pass in the two provinces are not different is tested against the hypothesis that the proportion of learners who passed to those who did not pass in the two provinces are significantly different after adjusting for both school quintile and gender.

The null (H_0) and alternative (H_1) hypotheses may be stated as:

H_0: $\hat{p}_{GP} = \hat{p}_{WCP}$ after adjusting for both school quintile and gender, and

H_1: $\hat{p}_{GP} \neq \hat{p}_{WCP}$ after adjusting for both school quintile and gender.

Table 4.25 indicates that learners from Western Cape Province than learners from Western Cape Province were about 1.51 more likely to pass Grade 12 in 2009 after adjusting for Quintile and Gender ($p < 0.01$, $OR = 1.507618$, 95% $CI : 1.466547 - 1.549838$).

In another words, the odds of passing Grade 12 increased by a factor of about 1.51 for learners from Western Cape Province over that of learners from Gauteng Province. We note that Quintile is a significant predictor of pass or not pass ($p < 0.001$), while gender is not a significant predictor (p = 0.263).

Table 4.25: Comparison of Provinces : 2009
Adjusting for Quintile and Gender

final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
Werstern Cape						
province	1.507618	.0212455	29.13	0.000	1.466547	1.549838
gender	1.01388	.0124875	1.12	0.263	.9896975	1.038652
quintile_2	1.207417	.0286422	7.95	0.000	1.152564	1.26488
quintile_3	1.460517	.031393	17.62	0.000	1.400265	1.52336
quintile_4	1.86013	.0399781	28.88	0.000	1.783402	1.940159
quintile_5	7.750707	.1760974	90.13	0.000	7.413135	8.103652
cons	.9508886	.0192914	−2.48	0.013	.9138201	.9894607
Quintile_1 is a reference quintile						
Gauteng Province is a reference province						

The null hypothesis is rejected, and we conclude that the proportions in the two provinces are significantly different. Therefore, learners from Western Cape Province than learners from Gauteng Province were more likely to pass Grade 12 after adjusting for both gender and school quintile.

Summary

Learners from Western Cape Province than learners from Gauteng Province were more likely to pass Grade 12 in 2009. All adjusted models indicate that the chance for learners from Western Cape Province increased by a factor of about 1.50 over that of Gauteng Province.

Precisely, if all learners belonged to the same gender-category (i.e., all learners were males or all learners were females) and all schools were equally resourced then the odds of passing Grade 12 would favour learners from Western Cape Province than learners from Gauteng Province. Therefore, the evidence to support the fact that learners from Western Cape Province would have better chances of passing Grade 12 than learners from Gauteng Province is sufficient.

c) **Adjusted models for 2010**

This section presents the adjusted models as fitted to the 2010 dataset. As with adjusted models for 2008 and 2009, Quintile, Gender, and then both Quintile and Gender will be adjusted when determining the estimates.

i) **Adjusting for Quintile**

We test the hypothesis that after adjusting for school quintile - the proportion of learners who passed Grade 12 to the proportion of those who did not pass Grade 12 in the two provinces are not different (H_0) againt the hypothesis that the proportion of learners who passed Grade 12 to the proportion of learners did not pass Grade 12 in the two provinces are significantly different (H_1).

We may state the null hypothesis and alternative hypothesis as follows:

H_0: $\hat{p}_{GP} = \hat{p}_{WCP}$ after adjusting for school quintile , and

H_1: $\hat{p}_{GP} \neq \hat{p}_{WCP}$ after adjusting both school quintile.

The odds of passing Grade 12 in 2010 favoured learners from Western Cape Province over learners from Gauteng Province. As shown in Table 4.26, after adjusting for Quintile, learners from Western Cape Province than learners from Gauteng were about 1.1 more likely to pass Grade 12 ($p < 0.01$, $OR = 1.090165$, 95% CI: $1.060064 - 1.12112$). Thus, the odds of passing Grade 12 increased by a factor of about 1.1 for learners from Western Cape Province over that of learners from Gauteng Province.

The hypothesis of no difference in proportions between the two provinces

Table 4.26: Comparison of Provinces : 2010
Adjusting for Quintile

final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
Western Cape						
Province	1.090165	0.0155739	6.04	0.000	1.060064	1.121120
quintile_2	1.145699	0.0276478	5.64	0.000	1.092772	1.201190
quintile_3	1.373247	0.0301596	14.44	0.000	1.315389	1.433649
quintile_4	1.749544	0.0383364	25.53	0.000	1.675996	1.826318
quintile_5	5.890063	0.1346791	77.55	0.000	5.631924	6.160034
cons	1.486703	0.0284356	20.73	0.000	1.432002	1.543494

Quintile_1 is a reference quintile

Gauteng Province is a reference province

isrejected. Therefore, we have sufficient evidence to conclude that learners from Western Cape Province were more likely than learners from Gauteng Province to pass Grade 12 after adjusting for school quintile.

ii) **Adjusting for Gender**

We test the hypothesis that after adjusting for gender - the proportion of learners who passed Grade 12 to the proportion of those who did not pass Grade 12 in the two provinces are not different (H_0) against the hypothesis that the proportion of learners who passed Grade 12 to the proportion of learners did not pass Grade 12 in the two provinces are significantly different (H_1).

We may state the null hypothesis and alternative hypothesis as follows:

H_0: $\hat{p}_{GP} = \hat{p}_{WCP}$ after adjusting for gender and
H_1: $\hat{p}_{GP} \neq \hat{p}_{WCP}$ after adjusting for gender.

Table 4.27 presents the results of the adjusted model (after adjusting for Gender). It can be seen that learners from Western Cape Province than learners from Gauteng Province were 0.87 less likely to pass Grade 12 ($p < 0.01$, $OR = 0.8749709, 95\% \; CI : 0.852898 - 0.8976151$). Clearly, the odds of passing Grade 12 is about 13% significantly lower for learners from Western Cape Province as compared to those for learners from Gauteng Province. Gender is a sigificant predictor of the binary outcome ($p = 0.006$).

We fail to accept the null hypothesis and we conclude that learners in the Western Cape Province were less likely than learners from

Table 4.27: Comparison of Provinces : 2010
Adjusting for Gender

final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]	
Western Cape						
Province	.8749709	.0114064	−10.25	0.000	.852898	.8976151
Gender	.9663287	.0120824	−2.74	0.006	.9429355	.9903023
cons	3.396634	.0351391	118.20	0.000	3.328456	3.466208
Gauteng Province is a reference province						

Gauteng Province to pass Grade 12 after adjusting for gender in 2010.

iii) **Adjusting for Quintile and Gender**

The null hypothesis that the proportions in the two provinces are not different is tested against the alternative hypothesis that the proportions in the two provinces are significantly different after controlling for both gender and school quintile.

Table 4.28 indicates that after adjusting for both school quintile and gender, learners from Western Cape Province than learners from Gauteng Province were about 1.10 more likely to pass Grade 12 in 2010 ($p < 0.01$, $OR = 1.09095$, $95\%\ CI : 1.060811 - 1.121946$). In other words, the odds of passing Grade 12 is significantly 10% more favouring learners from the Western Cape Province than for learners in Gauteng Province.

Table 4.28: Comparison of Provinces : 2010
Adjusting for Quintile and Gender

final	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]	
Western Cape						
Province	1.090950	.0155938	6.09	0.000	1.060811	1.121946
Gender	1.000559	.0130953	0.04	0.966	.9752185	1.026557
quintile_2	1.146124	.0276609	5.65	0.000	1.093172	1.201641
quintile_3	1.374181	.0301852	14.47	0.000	1.316275	1.434635
quintile_4	1.751001	.0383735	25.56	0.000	1.677383	1.827851
quintile_5	5.898432	.1349267	77.58	0.000	5.639821	6.168901
cons	1.485563	.0304666	19.30	0.000	1.427034	1.546493
Quintile_1 is a reference quintile						
Gauteng Province is a reference province						

There is sufficient evidence to conclude that after adjusting for both gender and school quintile; learners from Western Cape Province than learners from Gauteng Province were more likely to pass Grade 12 in 2010 academic year.

Summary

Learners from Western Cape Province than learners from Gauteng Province had better chances of passing Grade 12. The results from adjusted models indicated that the Western Cape Province learners than the Gauteng Province learners were significantly more likely to pass Grade 12 in 2010, except when the adjustment was made for gender only.

Chapter Summary

The adjusted and unadjusted causal models were fitted for 2008, 2009 and 2010 datasets. The results favoured the Western Cape Province over Gauteng Province. In other words, when we adjusted for quintile (and/or gender) the odds of passing Grade 12 favoured the Western Cape Province than Gauteng Province. However, adjusting for gender only shows that the odds of passing Grade 12 favoured Gauteng Province only if the outcome was binary and the academic year was 2010. Thus, if learners in the two provinces were all of the same gender-group then learners from Gauteng Province would have had better chances of passing Grade 12 than learners from Western Cape Province in 2010 academic year. In the event that all schools in the two provinces were equally resourced (and that all learners belonged to the same gender-group) then the odds of passing Grade 12 favoured learners from Western Cape Province.

Chapter 5

Conclusion and Recommendations

*****

"In a coeducational school, boys and girls learn together, converse together and grow into adulthood together. They're at ease with one another and, in my personal experience, more at ease with themselves."

http://www.telegraph.co.uk/education/educationopinion/11476686/Are-single-sex-or-mixed-schools-the-way-forward.html

Chapter Preview

This chapter summarises the major findings of this study and outlines some recommendations. Section 5.1 emphasises the importance of utilising statistical sofware when dealing with clustered data since clustered data require computer-intensive methods. Section 5.2 highlights the importance of collecting all significant covariates when forming or creating database for learners. Section 5.3 summarises the results in terms of the effects of all available covariates. Finally, Section 5.4 makes some general recommendations.

5.1 Statistical Models

Regression analyses, in particular multiple linear regression (MLR) analyses [see Chapter 3, Page 64], are commonly employed in social sciences (Montgomery and Peck, 1982; Fox, 1997; Weisberg, 2005). These models are extension of simple linear regression models. MLRs are popular because they allow an investigator, researcher or analyst to answer questions that consider the role(s) that multiple independent covariates play in accounting for variance in a single predictor variable (Montgomery and Peck, 1982; Weisberg, 2005, Page 47). We note that aggregated analyses and/or descriptive analyses are unable to adjust for individual covariates (Donner and Klar, 2000). The drawback of MLR methods is the fact that they are unable to adjust or account for ICC.

The aim of the study was to highlight the methods for analysing clustered data in nonrandomisation studies. The aim was achieved with the application of hierarchical models, also called causal models. The researcher applied these models to the Grade 12 data as supplied by Umalusi. Grade 12 data is a typical clustered data since it contains data on learners' results per learning subject (area). Thus, learners were observed more than once even though this study is cross-sectional (see hypothetical examples as depicted by Figure 3.2 on Page 76 and Figure 3.3 on Page 77).

With the development of (new) statistical techniques, such as GEE and random coefficient analysis, it has become possible to analyse clustered or longitudinal relationships (Skrondal and Rabe-Hesketh, 2004) using all available clustered data, without performing data resolution (Twisk, 2003; Letsoalo and Lesaoana, 2012).

Popular methods for analysis of binary response data include the probit model, discriminant analysis, and logistic regression (Hailpern and Visintainer, 2000; Hilbe, 2009). Logistic regression is the method of choice since it makes no assumption about the variable distribution, and it is a direct probability model because it is stated in terms of $Pr\{Y = 1|X\}$ (Hosmer and Lemeshow, 2000; Hilbe, 2009). Its added advantage is its ability to provide valid estimates, regardless of study design (Hailpern and Visintainer, 2000). However, caution must be adhered to when observations of interest are not independent (Letsoalo and Lesaoana, 2012).

Correlated or clustered data are common in social sciences especially in observational studies. Standard statistical methods, such as student's t-test and classical regression models are inappropriate for clustered data due to their assumption of independence between study units (Skrondal and Rabe-Hesketh, 2004; Letsoalo and Lesaoana, 2012).

This study took note of the existence of ICC in the analysis of the data. It was not assumed that $ICC = 0$ since that would have biased the point estimates, hence leading to falsely narrow confidence intervals (Osborne, 2000; Twisk, 2003; Wu and Zhang, 2006; Gelman and Hill, 2007; Letsoalo and Lesaoana, 2012).

Cluster randomised studies are more applicable in controlled settings. There are two main reasons to randomise at a level larger than the individual. First, it can address contamination: where treated individuals mix and chat and potentially "share" treatment with individuals in the comparator group. This would "contaminate" the impact, and comparator group would no longer be a good comparison. Randomising at the higher level may minimise the risk of this happening. Second, randomisation might be applied at the intervention level, i.e. the level that the intervention would actually be implemented.

When randomising at the cluster level, the unit of randomisation is the unit at which we will randomly roll out the program or intervention. The unit of analysis, defined as the unit at which we will collect data and compare outcomes, is usually the individual - for example, individual learners' test or examination scores. This distinction is crucial when determing sample size. Among other things, sample size is affected by ICC, which refers to how similar or dissimilar individuals within a cluster are (Donner and Klar, 2000; Jonh, 2013). Application of standard sample size approaches to cluster randomisation trials or designs may lead to an underpowered study (Type *II* error), and on ther other hand the application of standard statistical methods to clustered data generally tends to bias *p*-values downwards, i.e. could lead to spurious statistical significance (Type *I* error) (Donner, 1998; Donner and Klar, 2000; Jonh, 2013).

5.2 The Grade 12 Data

The Grade 12 dataset did not have all predictor variables such as parents' attributes, family attributes, learners' background and learners' age. From other studies, see Chapter 1, we know that these covariates significantly predict learners' performance. It would have been very useful to test if these variables would yield similar results in the South African setting. In particular, we could not test for the significance of these covariates since they were not contained in our dataset (see Appendix A, Section A3, Page 142). It may be of great importance that Umalusi collect all (possible) covariates such as the gender of educators, parent's SES, qualifications of educators, education level of parents and so on, for further studies and analyses.

5.3 Results

This section comments on the findings in this study. The discussion of results is given in two parts; namely the discussion of descriptive statistics and discussion of estimates as given by the fitted causal models. The estimates are generated from the unadjusted models and the adjusted models.

5.3.1 Descriptive Statistics

The use of descriptive statistics assists the researchers with identification of measures of cental tendencies and summary statistics. Descriptive statistics gives the so-called "overall picture" with regard to the data distribution. The descriptive analysis' results show that the pass rates favoured Western Cape Province in 2008 (77.08% vs. 74.33%) and 2009 (73.96% vs. 70.42%) academic years. Otherwise the pass rates favoured Gauteng Province only in 2010 (76.90% vs. 74.43%).

5.3.2 Statistical Estimates

The results from the both crude and adjusted models gave a better insight. The presented results showed the effect of gender, school quintile, and gender and quintile. Essentially, Western Cape Province performed better than Gauteng Province upon adjustment for these covariates, with the exception of the academic year 2010 when adjustment was made only for gender.

5.3.2.1 Effect of Gender

The current practice emphasises the comparisons of the proportion of male learners who passed Grade 12 to the proportion of female learners who passed Grade 12. This study goes beyond just proportions. Through the use of causal models it has been shown that gender has a significant effect on overall learner performance. That is, with this insight the department of education can factor in gender in its intervention strategies.

The models for predicting final marks (%) indicated that the two provinces performed significantly differently after controlling for gender. Also, the models indicate that gender is a significant predictor of overall performances of learners in the two provinces. Therefore, male learners performed significantly different from female learners in the two provinces ($p < 0.01$).

This result suggests that single-sex schooling system may be encouraged in the two provinces. At single-sex schools, males and females can explore educational opportunities without being constrained by expectations, stereotypes (e.g. female learners are regarded as being weak in mathematics and science). For example, girls at single-sex schools are more likely to explore nontraditional subjects and are encouraged to be daring and invest in subjects they might otherwise not try if they were in mixed-sex schools.

Okoro et al. (2012) indicated that male and female learners do not seem to exhibit the same level of academic achievement. Also, as emphasised by Herr and Arms (2004) part of the rationale for single-sex schooling is the view that adolescents create a culture in school that is at odds with academic performance and achievement. In coeducational settings, the culture is one of socialisation where for some, academics might not be a priority. For these individuals, single-sex classes or single-sex schools might be a better choice. Also, with single-sex classes or schooling, the more targeted teaching is possible.

The same cannot be said in cases where the outcome is binary (pass or not pass). The results indicate that gender is not a significant predictor of whether or not a learner will pass Grade 12 in the two provinces. Therefore, the single-sex mode of schooling cannot be encouraged in the two provinces especially when the interest of the authorities is binary outcome. Generally speaking, the main claims for co-education are to do with personal and social development. It has been argued that the sexes growing up together, including in school, brings a number of benefits including:

1) greater happiness, better behaviour and fuller emotional development;

2) smoother transition to the mixed environment of university and life generally;

3) and hence, parents and pupils prefer it.

There is a confounding factor that can be thought of as highly significant contributor in determining whether or not a learner can perform better, and that factor is the ability of a learner. Therefore learner ability and learner background are thought of as essential and should be taken into account in province, region or school comparisons. If a fitted model can adjust for ability, SES, school quintile, learners' first language, settlement type (rural, semi-rural, urban, etc) and ethnicity then apparent advantages to single-sex or co-education can emerge. It is difficult to prove that the so-called good girls' school or good boys' school are good enough because are single-sex. On the other hand, it is equally difficult to prove that the so-called bad boys' schools and bad girls' schools are so because they are single-sex schools.

> *"Although single-sex is often treated as one category, it cannot be assumed that all-girls and all-boys schools are the same. If it is true, as it is sometimes claimed, that girls can achieve more in their own schools because the boys are not there to dominate, disrupt or distract, one has to wonder about the effects of doubling the number of disruptors and distracters as would be the case in all-boys schools - unless it is further claimed that this behaviour is only elicited in the presence of girls. You cannot have an educational system in which single-sex education is just for the one sex."*
>
> *Unknown Source*

> *"As educators, and as people, we tend to assume that females and males are different — are indeed "opposite sexes." We see someone's sex as an important predictor of their abilities and interests and assume that if we know someone is a girl or a boy, we know a lot about them. That assumption is wrong! Knowing someone's sex may tell us a lot about them biologically but it tells us very little about them in other ways."*
>
> *Unknown Source*

5.3.2.2 Effect of Quintile

The findings in this thesis indicate that the two provinces performed significantly different after adjusting for quintile. Also, quintile was found to be a significant predictor of the final mark (%) in the two provinces. In particular, better resourced schools performed better than less resourced schools. This may suggests that the-how-resourceful the school is might determine how the learners will perform in the provinces.

In case where the outcome is dichotomous (*Pass/Not Pass*); the two provinces performed differently, and school quintile was found to be significantly predicting the outcome. The resourceful schools are more likely than the less resourced schools to produce the *Pass* result. In another words, learners in the resourceful schools are more likely than learners in the less resourced schools to pass Grade 12.

In the South African context, the less resourced schools are more likely to be in rural areas or townships. Schools in rural areas lack infrastructure, quality educators, and parents in such areas are more likely to be "uneducated" or less educated. The education of parents plays a significant role in the learning process of their children. More so that the education level of parents is tightly linked to the SES of the parents (and or family). The relationship between SES and learner performance was found to be positive (Gnanamoorthy, 2014). Simply put - learners who come from high SES families than those who come from low SES are more likely to perform better. Mayer (2002) states that poor learners have less chance of performing well or have more chances of performing poorly. Better performance may mean obtaining higher marks (%) or obtaining *pass* results rather than 'not pass' results.

5.3.2.3 Effect of Gender and Quintile

The Western Cape Province performed significantly better than Gauteng Province after adjusting for gender and quintile in 2008, 2009 and 2010 academic years. This statement is valid for the continuous endpoint (marks [%]) . However, it is worth noting that both gender and quintile are significant predictors of the overall learner performances. If the authorities are interested in the overall marks (%) then a single-sex classroom or single-sex schooling with resourced schools is encouraged.

Gender is not a significant predictor of the binary endpoint (pass or not pass) after adjusting for Province and Quintile, e.g., Tables 4.22 and 4.22. Therefore, a mixed-sex schooling is favoured if the stakeholders are interested only in the binary outcome (pass or not pass) in the two provinces.

The results in this thesis do not only indicate the difference between the two provinces but quantify these differences. This (observation or) conclusion cannot be achieved by the use of descriptive statistics since it (descriptive statistics) concerns itself only with the measure of central tendencies and proportions (Donner and Klar, 2000; Osborne, 2000; Gelman and Hill, 2007; Hayes and Moulton, 2009; Nathans et al., 2012).

5.3.2.4 Hypothesis

The hypothesis that the proportions in two study provinces were not significantly different was not accepted. Also, the hypothesis that the average performances in the two provinces were the same was not acceptated. In another words, it has been established through the use of causal models that there was a sufficient evidence to conclude that the two study provinces were significantly different.

5.4 Recommendations

5.4.1 Statistical Models

The results of the study have shown that causal models are plausible in the analysis of clustered data, and that these models must be adopted as the tools for analysis of Grade 12 results or in the investigation of school effectiveness, and hence their performances. With the application of these models; the researcher took heed of the correlatedness or dependencies in the data. Also, these models are able to adjust for other covariates. Hence, dealing with confounding effects.

The argument in this thesis is that the use of summary statistics is flawed - a necessary exercise in self-deception. Causal models provide schools with good quality comparative information (interpretation of results) about their effectiveness and performances; for the effectiveness of schools will provide the Department of

Basic Education and other relevant stakeholders with the knowledge of how and when to implement their interventions.

To be precise, causal models - if used to analyse clustered data such as Grade 12 data - enable researchers to identify the highly significant or important predictor variables over a period of time. As such, these models are helpful since stakeholders will be in a know of the important covariates and be able to put suitable intervention strategies in place. In other words, causal models may be found to be helpful in terms of long term planning since they are able to quantify and identify significant predictors of outcome variables.

5.4.2 Creation of Database

The creation of database with all covariates which are suspected to be significantly associated with outcome, e.g. SES, learner ability and educator's gender are believed to be associated with learner performance, would be helpful. To be precise, the recommendation is that based on the findings from different studies (although mixed); Umalusi is encouraged to start collecting and creating database of learners' attributes (e.g. family SES, mothers' education, parents' occupation etc.) and schools attributes (e.g. school quintile, location, teacher gender, teacher quilification and so on.); for it is desirable to test (for the effect of) these variables in South African setting.

5.4.3 Summary Statistics

Descriptive statistics or summary statistics offers no way of justifying if the (observed) marginal differences between the study arms are significant. Strictly speaking, descriptive statistics offers no (sufficient) evidence to conclude whether or not the difference one observes is significant. For example, in this study we have no (sufficient) evidence that the (observed) marginal differences, 1.75%, 3.54% and 2.47%, are significant. Therefore, descriptive statistics should not be treated as the beginning and the end; for the researchers are more likely to commit an error in reasoning, especially if extrapulations are to be made based on the results.

5.4.4 Estimates

The effects of all covariates contained in the Grade 12 dataset were examined through the adjusted models and unadjusted models.

5.4.4.1 Gender

The effect of gender gave mixed results in that when the outcome of interest was categorical (binary) then the effect of gender did not significantly predict the outcome. However gender significantly predict the overall total marks.

In their planning and intervention strategies - it can be explicitly mentioned that authorities and (or) stakeholders need to be precise with the objective of the intervention; for:

a) if the outcome is continuous marks then the issue of learner gender becomes crucial. That is, if the authorities are interested in producing the learners who will score relatively higher marks (say), then they (authorities) should factor in the issue of learner gender. For example, male learners perform significantly better than when they are being taught by female educator. Similarly, female learners perform better when they are being taught by a female educator than when they are being taught by male educator [for example, see Okoro and Uwah (2013)]. Therefore, learner gender becomes a significant pointer when the outcome is continuous or when authorities are interested in relatively higher scores.

b) if the outcome is binary (pass or otherwise) then learner gender is not a significant issue. In another words, though learner gender is important, it is not a significant predictor of the outcome. Even though it could not be proven statistically that learner gender predicts whether a learner will pass or not pass Grade 12 in the two study provinces, learner gender remains an important covariate to consider (at least) from social point of view.

The employ of causal model on clustered Grade 12 data revealed that learner gender is crucial in the planning of the interventions by the stakeholders. The intervention strategies cannot afford to ignore the effect of gender especially when the authorities' interest is on learners achieving or scoring higher overall marks.

5.4.4.2 School Quintile

It is common knowledge that better resourced schools have the ability to produce better performing learners. This has been captured through the comparisons of proportions between the school quintiles. The application of causal models provided sufficient evidence of the effect of school quintile.

School quintile significantly predicts the outcome irrespective of the nature of the outcome, binary or continuous. Since less resourced schools are more likely to be in poor communities or rural settlements - the authorities may be advised to redistribute resources in such a way that rural schools attract quality educators by introducing the so-called rural incentives for educators, and start investing in the infrastructure in those areas.

We note that resources are not directly linked or associated with improvement of learning outcome; for resources may be necessary but they are not sufficient. Differences in their effects depend on differences in their use Kurdziolek (2011).

Stakeholders (authorities, educators and community at large) in under-resourced schools must engage in school improvement plan (SIP) so that these schools progress towards whole-school development; for if stakeholders do not get involved in the planning process, and the SIP[1] does not contain specific and measurable targets and there is no proper analysis of educators' and learners' performance and attendance and the monitoring of these issues is not systematic then these schools (under-resourced) will remain poor-performing schools (Van Der Voort and Wood, 2014). It is noted, as highlighted by Van der Berg (2008), that more resources did not necessarily or without qualification improve school performance, although some resources (e.g. equipment at the school) appeared to play a role. As in much of the educational production function literature, the message from Van der Berg (2008) appeared to be not that resources did not matter, but rather that resources mattered only conditionally.

The vast majority of mainly black african learners enrolling at secondary or high schools come from rural areas or township school where a lack of resources and educator training create an environment of rote learning where learners leave with only a superficial understanding of some of the linguistic and numeracy concepts needed to successfully pass Grade 12.

5.4.5 Design Issues

The effect of education on learners' performance has proven to have suffered severely due to clustered nature of educational data. An attempt to perform aggregated data analysis has proved to be misleading [as sample size is reduced - leading to loss of information (Letsoalo and Lesaoana, 2010)]. This being the case, the effects of education can exist both between and within the units at each level of the educational system; the majority of studies of educational effects have restricted attention to either overall between-student, between-class, or between-school analyses. Therefore, the majority of studies, of educational effects carried out conceal more than they reveal, and as such ordinary regression methods and descriptive statistics reported have

[1] "The school improvement plan forms the basis for the continuous school improvement, as well as acting as a monitoring instrument to measure progress towards specific areas of whole-school development" (Van Der Voort and Wood, 2014).

most likely generated unreliable conclusions in many studies wherein the analysed data was clustered.

5.4.6 Further Studies

Statistical analysis is a foundation of all controlled studies e.g. biological research. Various methods exist to test whether there is a difference in the response between study arms. Observations are generally made across a series of independent experiments to ensure that a result is reproducible. It is noted that results from a single experiment lack reproducibility. Statistically significant differences between groups (typically, p < 0.05) indicate that the observed difference is unlikely to have occurred by chance, suggesting a “real” difference between groups. Research outcomes rely on the presentation of statistically valid conclusions, and thus the approach used for statistical analysis is critical.

Clustered data arise when the data from the whole study can be classified into a number of different groups, referred to as clusters. Each cluster contains multiple observations, giving the data a “nested” or “hierarchical” structure, with individual observations nested within the cluster. The key feature of clustered data is that observations within a cluster are “more alike” than observations from different clusters (Hox, 1995; Sullivan et al., 1999; Donner and Klar, 2000; Gelman and Hill, 2007; Letsoalo and Lesaoana, 2010, 2012).

Although social sciences are mainly characterised by nonrandomisation designs - causal model (applied with propensity score technique) may find its application in educational setting wherein the *effectiveness of the intervention* is to be measured. Here, one may need to determine the effect of matric results in the performance of university students in the junior degree programmes; for student will be observed at several time points or the observed data will be clustered. Therefore, causal models may be applicable in the areas of monitoring and evaluation.

Also, causal model may find its application in situations where a researcher is to *validate* the effect or impact of the intervention. For example, in a counterbalancing design or crossover design, a repeated measurements design such that each experimental unit (learner) receives different treatments during the different time periods, i.e., the learner cross over from one teaching and learning approach (A) to another (B) during the course of the trial. In another words, experimental design in which all possible orders of presenting the variables are included. For example, if you have two groups of participants (group 1 and group 2) with each member of the group being observed at least once, and two levels of an independent variable, Level a and Level b (Cozby, 2009). The objective of counterbalancing design is to determine whether the two interventions yield equivalent effects (e.g. bioequivalence trials).

The reason to consider a counterbalancing design when planning a social science study is that it could yield a more efficient comparison of treatments than a parallel design, i.e., fewer participants might be required in the counterbalancing design in order to attain the same level of statistical power or precision as a parallel design. Intuitively, this seems reasonable because each participant or learner serves as his/her own matched control (Cozby, 2009). Causal models may be plausible if each participant was observed more than once under each treatment condition. Therefore, the ICC between the observation for individual participant is more likely to exceed zero - justifying the use of HLMs (in particular, the causal models).

The order in which treatments or interventions are given can actually affect the behaviour of the participants or elicit a false response, due to fatigue or outside factors changing the behaviour of many of the participants. To counteract this, researchers often use a counterbalanced design, which reduces the chances of the order of treatment or other factors adversely influencing the results (Cozby, 2009); for counterbalancing design or crossover design is mainly a controlled design.

Chapter Summary

The importance of fitting hierarchical models or causal models (disaggregated analysis) to clustered data was emphasised. It was noted that causal models are plausible for long term planning as they are able to determine the estimate of the effect of individual covariates. All important covariates were discussed such that effect of gender gave rise to controversial social issues. While it may be suggested that single-sex schooling can be encouraged- the argument could not be sustained over mixed-sex schooling. These findings will always remain elusive if one uses aggregated approach to data analysis - in particular, descriptive statistics. Causal models may yield reasonable results in the areas of validating the approaches and in the areas wherein a researcher is to measure the effect of the intervention (monitoring process).

Appendix A

Grade 12 Dataset, and Stata Commands and Summary Statistics

*****

A1 Descriptive Statistics

a) Categorical Variables (gender and final)

- *bys province year: xttab gender, i(candidate)*
  generates Table 4.1 on page 96.

- *bys province year: xttab final, i(candidate)*
  generates Table 4.2 on page 96.
- *bys year: tab province assess if identifier==1, chi2 row*
  generated Table 4.4 on Page 101.

b) Continuous Variable

- *bys province year : xtsum marks, i(candidate)*
  generates Table 4.3 on page 97.

A2 Hierachical Models

a) **Categorical Outcome [Binary (Pass or Not Pass)]**

- *bys province year: xi: xtlogit final i.province [iweight = weight], i(candidate) pa*
- *bys province year: xi: xtlogit final i.province i.gender [iweight = weight], i(candidate) pa*

- *bys province year: xi: xtlogit final i.province i.quintile [iweight = weight], i(candidate) pa*

b) Continuous Outcome [marks (%)]

- *bys province year: xi: xtreg marks i.province [iweight = weight], i(candidate) mle*
- *bys province year: xi: xtreg marks i.province i.gender [iweight = weight], i(candidate) mle*
- *bys province year: xi: xtreg marks i.province i.gender i.quintile [iweight = weight], i(candidate) mle*
- *bys province year: xi: xtreg marks i.province i.quintile [iweight = weight], i(candidate) mle*

A3 Dataset: Data Dictionary

The dataset contains variables:

a) School Number (**school**) which is a dummy school identifier,

b) Centre Type (**centre**) that idicates whether a school was a public, independent or special,

c) Quintile (**quintile**) - indicates the resourcefulness of the school (1 stands for least resourced and 5 stands for mostly resourced),

d) Candidate-Exam-Number (**candidate**) - dummy learner number,

e) Exam-Date (**year**) - indicates the year in which a learner wrote Grade 12 examination,

f) Gender (**gender**) which indicates whether a learner was a male or female

g) Subject-Code (**subject**) is the abbreviation/accronym for subject name

h) Percentage (**marks**) - the final mark (%) obtained by a learner in a particular learning subject,

i) Higher-Education-Assessment (**assess**) is the final outcome or recommendation that indicates the overall learner achievement, and

j) Province (**province**) which indicates the province in which a learner wrote grafe 12 examination.

The binary variable **final** that indicates whether or not a learner passed Grade 12 was generated from the variable **assess**.

A4 Summary Statistics

Tables A.1 and A.2 give the detailed distributions of learners in the two provinces according to gender and pass rates for academic years 2008, 2009 and 2010, respectively. The tables indicated the occasion or repetition within each learner.

Table A.1: Distributions of gender by Provinces: 2008, 2009 and 2010 academic years.
Note: *Overall* indicates the frequency that includes replicates or duplicates and *Between* is the count that excludes repeatition. Table 4.1 is a portion of this table.

	Gauteng	Overall		Between		**W. Cape**	Overall		Between	
		Freq.	Percent	Freq.	Percent		Freq.	Percent	Freq.	Percent
2008	Male	313796	45.48	44940	45.44	Male	134738	42.91	19268	42.84
	Female	376166	54.52	53954	54.56	Female	179254	57.09	25709	57.16
	Total	**689962**	**100.00**	**98894**	**100.00**	**Total**	**313992**	**100.00**	**44977**	**100.00**
		$(n = 98894)$					$(n = 44977)$			
2009	Male	317163	45.58	45315	45.52	Male	139375	43.16	19930	43.11
	Female	378640	54.42	54234	54.48	Female	183550	56.84	26304	56.89
	Total	**695803**	**100.00**	**99549**	**100.00**	**Total**	**46234**	**100.00**	**46234**	**100.00**
		$(n = 99549)$					$(n = 46234)$			
2010	Male	297274	45.09	42475	45.01	Male	141393	43.30	20232	43.22
	Female	361997	54.91	51893	54.99	Female	185153	56.70	26584	56.78
	Total	**659271**	**100.00**	**94368**	**100.00**	**Total**	**326546**	**100.00**	**46816**	**100.00**
		$(n = 94368)$					$(n = 46816)$			

Table A.2: Distributions of Pass rates by Provinces: 2008, 2009 and 2010 academic years.
Note: *Overall* indicates the frequency that includes repeatitions and *Between* is the count that excludes repeatition.
Table 4.2 was generated from this table.

	Gauteng	Overall		Between		**W. Cape**	Overall		Between	
		Freq.	Percent	Freq.	Percent		Freq.	Percent	Freq.	Percent
2008	Fail	170797	24.75	25381	25.67	Fail	69190	22.04	10309	22.92
	Pass	519165	75.25	73512	74.33	Pass	244802	77.96	34668	77.08
	Total	**689962**	**100.00**	**98893**	**100.00**	**Total**	**313992**	**100.00**	**44977**	**100.00**
			$(n = 98893)$					$(n = 44977)$		
2009	Fail	199779	28.71	29443	29.58	Fail	81252	25.16	12040	26.04
	Pass	496024	71.29	70106	70.42	Pass	241673	74.84	34194	73.96
	Total	**695803**	**100.00**	**99549**	**100.00**	**Total**	**322925**	**100.00**	**46234**	**100.00**
			$(n = 99549)$					$(n = 46234)$		
2010	Fail	145342	22.04	21797	23.10	Fail	80411	24.62	11974	25.57
	Pass	513982	77.96	72580	76.90	Pass	246221	75.38	34856	74.43
	Total	**659324**	**100.00**	**94377**	**100.00**	**Total**	**326632**	**100.00**	**46830**	**100.00**
			$(n = 94377)$					$(n = 46830)$		

Appendix B

Logistic Regression Model

*****

In linear regression with continuous outcome variable y_i, a model of the form

$$E(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}, i = 1, 2, \cdots, n \tag{B0.1}$$

could be fitted. However, one encounters a problem with expressing $E(y_i) = \pi_i$ where π is the probability (Binomial distribution), and $0 \leq \pi_i \leq 1$. Employing a tranformation function g of π_i that maps $(0, 1)$ onto $(-\infty, \infty)$ then $g(\pi_i$ is expressed as a linear combination of the predictor variables

$$g(\pi_i) = \beta_0 + \sum_{j=1}^{p} x_{ji}\beta_j, i = 1, 2, \cdots, n \tag{B0.2}$$

One such transformation is the logit,

$$g(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i} = \beta_0 + \sum_{j=1}^{p} x_{ji}\beta_j, i = 1, 2, \cdots, n \tag{B0.3}$$

$$= \eta_i, \tag{B0.4}$$

B1 Estimation

Detailed works on estimation procedures are given by, among others, McCullagh and Nelder (1989) and Hardin and Hilbe (2001).

B11 Likelihood function

The coefficients β_0 to β_p can be estimated by maximising the likelihood function. For individuals with $y_i = 0$, the contribution to the likelihood is $(1 - \pi_i)^{1-y_i}$. If $y_i = 1$ then the contribution is $\pi_i^{y_i}$. The contribution of any observation is thus $\pi_i^{y_i}(1 - \pi_i)^{1-y_i}$. The likelihood function for n observations is thus

$$L(\beta)) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, i = 1, 2, \cdots, n. \tag{B1.1}$$

The likehood depends on the π_i which in turn depends on β.

Maximum likelihood estimates $\hat{\beta}_i$ are obtained by finding the values which maximise L, or equivalently $\ell n L$.

$$\begin{aligned}
\ell n L(\beta) &= \sum_{i=1}^{n}[y_i \ell n \pi_i + (1-y_i)\ell n(1-\pi_i)] \\
&= \sum_{i=1}^{n}\left[y_i \ell n \frac{\pi_i}{1-\pi_i} + \ell n(1-\pi_i)\right] \\
&= \sum_{i=1}^{n}\left[y_i \eta_i + \ell n\left(1 - \frac{e^{\eta_i}}{1+e^{\eta_i}}\right)\right] \\
&= \sum_{i=1}^{n}\left[y_i \eta_i - \ell n\left(1+e^{\eta_i}\right)\right]
\end{aligned}$$

where

$$\eta_i = \ell n \frac{\pi_i}{1-\pi_i} = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ji}. \tag{B1.2}$$

Taking partial derivatives

$$\begin{aligned}
\frac{\partial \ell n L}{\partial \beta_0} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{\eta_i}(1+e^{\eta_i})^{-1} \\
\frac{\partial \ell n L}{\partial \beta_j} &= \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} e^{\eta_i}(1+e^{\eta_i})^{-1} x_{ji}, j = 1, \cdots, p
\end{aligned}$$

Equatinf these $p+1$ equations to zero gives a set of $p+1$ non-linear equations which have to be solved.

B11 Newton-Raphson Procedure

The score of the j^{th} parameter is $\partial \ell n L / \partial \beta_j$. Denote the $(p+1) \times 1$ vector of scores by $U(\beta)$. A $(p+1) \times (p+1)$ matrix of secons order partial derivatives can be formed with (i, j)-j^{th} element

$$\frac{\partial^2 \ell n L}{\partial \beta_i \partial \beta_j}, i = 0, \cdots, p; j = 0, \cdots, p \tag{B1.3}$$

This matrix is called the Hessian matrix, denoted by $H(\beta)$. Near $\hat{\beta}$, at β_m, the Taylor's expansion of the scores vector gives

$$U(\hat{\beta}) \approx U(\beta_m) + H(\beta_m)(\hat{\beta} - \beta_m). \tag{B1.4}$$

The maximum likelihood estimates of the β's must satisfy

$$\frac{\partial \ell n L}{\partial \beta_j} = 0.$$

so

$$U(\hat{\beta}) = 0$$

and from B1.4

$$U(\beta_m) + H(\beta_m)(\hat{\beta} - \beta_m) = 0. \tag{B1.5}$$

Thus

$$\hat{\beta} \approx \beta_m - H^{-1}(\beta_m)U(\beta_m) \tag{B1.6}$$

which suggests an iterative scheme for estimating $\hat{\beta}$ such that

$$\hat{\beta}_{m+1} = \hat{\beta}_m - H^{-1}(\hat{\beta}_m)U(\hat{\beta}_m) \tag{B1.7}$$

B11 Fisher scoring

An alternative method of solving the likelihood equations is the Fisher scoring method. In here the Hessian matrix is replaced by the the matrix of expected values of second order partial derivatives. The information matrix I has $(i,k)^{\text{th}}$ element

$$-E\left[\frac{\partial^2 L}{\partial \beta_j \partial \beta_k}\right]. \tag{B1.8}$$

The iterative scheme is then

$$\hat{\beta}_{m+1} = \hat{\beta}_m + I^{-1}(\hat{\beta}_m)U(\hat{\beta}_m). \tag{B1.9}$$

In the case of linear logistic regression model $I^{-1}(\hat{\beta}) = -H^{-1}(\hat{\beta})$ so the two algorithms will not only converge to the maximum likelihood estimate of β, but will give the same standard errors of the parameter estimates (the square root of the diagonal elements of $-H^{-1}$ and I^{-1}).

B11 Iteratively Reweighted Least Squares

To fit linear logistic model using Fisher scoring, we need expressions for $U(\beta)$ and $I(\beta)$. From (B1.1) the log-likelihood function for n observations is given by

$$\ell nL = \sum_{i=1}^{n}[y_i \ell n\pi_i + (1-y_i)\ell n(1-\pi_i)] \tag{B1.10}$$

Thus

$$\frac{\partial \ell nL}{\partial \beta_j} = \sum_{i=1}^{n}\frac{\partial \ell nL}{\partial \pi_i}\frac{\partial \pi_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{B1.11}$$

where

$$\frac{\partial \ell nL}{\partial \pi_i} = \sum_{i=1}^{n}\frac{y_i}{\pi_i} - \frac{1-y_i}{1-\pi_i} = \sum_{i=1}^{n}\frac{y_i-\pi_i}{\pi_i(1-\pi_i)} \tag{B1.12}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ji} \tag{B1.13}$$

$$\frac{\partial \pi_i}{\partial \eta_i} = \frac{\partial \eta_i^{-1}}{\partial \pi_i} = g'(\pi_i) = \frac{1}{\pi_i(1-\pi_i)} \tag{B1.14}$$

Therefore, from (B1.15), (B1.13) and (B1.14)

$$\frac{\partial \ell nL}{\partial \beta_j} = \sum_{i=1}^{n}\frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{1}{g'(\pi_i)}x_{ji}. \tag{B1.15}$$

If $e_i = (y_i - \pi_i)g'(\pi_i)$ and

$$w_i = \frac{1}{\pi_i(1-\pi_i)[g'(\pi_i)]^2} = \pi_i(1-\pi_i) \tag{B1.16}$$

then

$$\frac{\partial \ell nL}{\partial \beta_j} = \sum x_{ji}w_ie_i \tag{B1.17}$$

therefore,

$$U(\beta) = X'We \tag{B1.18}$$

where X is the $n \times (p+1)$ matrix of p predictor variables plus the intercept term, W the $n \times n$ diagonal matrix of weights w_i, and the e is the $n \times 1$ vector with i-th component e_i.

To obtain the $(i,j)^{\text{th}}$ element of the information matrix we use

$$-E\left(\frac{\partial^2 \ell n L}{\partial\beta_j\partial\beta_k}\right) = E\left(\frac{\partial \ell n L}{\partial\beta_j}\frac{\partial \ell n L}{\partial\beta_k}\right). \tag{B1.19}$$

If $i \neq i'$ then

$$E[(y_i-\pi_i)(y_{i'}-\pi_{i'})] = Cov(y_i, y_{i'}) = 0. \tag{B1.20}$$

If $i = i'$ then

$$E[(y_i-\pi_i)^2] = Var(y_i) = \pi_i(1-\pi_i) \tag{B1.21}$$

Therefore,

$$\begin{aligned}
-E\left(\frac{\partial^2 \ell n L}{\partial\beta_j\partial\beta_k}\right) &= \sum_{i=1}^{n}\frac{\pi_i}{[\pi_i(1-\pi_i)]^2}\frac{(1-\pi_i)}{[g'(\pi_i)]^2}x_{ji}x_{ki}\\
&= \sum_{i=1}^{n}\frac{1}{\pi_i(1-\pi_i)}\frac{1}{[g'(\pi_i)]^2}x_{ji}x_{ki}\\
&= \sum_{i=1}^{n}x_{ji}w_i x_{ki}
\end{aligned}$$

so that

$$I(\beta) = X'WX. \tag{B1.22}$$

From (B1.9) $\hat\beta_{m+1} = \hat\beta_m + I^{-1}(\hat\beta_m)U(\hat\beta_m)$. Thus

$$\begin{aligned}
\hat\beta_{m+1} &= \hat\beta_m + (X'W_mX)^{-1}X'W_me_m\\
&= (X'W_mX)^{-1}[X'W_m(X\hat\beta_m + e_m)]\\
&= (X'W_mX)^{-1}X'W_m(\hat\eta + e_m)\\
&= (X'W_mX)^{-1}X'W_mz_m
\end{aligned}$$

where subscript m denotes the values obtained from the m^{th} iteration. $\hat\beta_{m+1}$ is thus obtained by regressing (using weighted least squares) the adjusted dependent variable z_m, with i^{th} element

$$\bar\eta_{im} + (y_i-\pi_i)g'(\pi_i) = \bar\eta_{im} + \frac{(y_i-\pi_i)}{\pi_i(1-\pi_i)} \tag{B1.23}$$

on the p predictor variables using weights w_{im} where

$$w_{im} = \frac{1}{\pi_i(1-\pi_i)\,[g'(\pi_i)]^2} = \pi_i(1-\pi_i) \tag{B1.24}$$

As initial estimates of π_i, $\bar{\pi}_{i0} = (y_i + 0.5)/2$ can be used, from which initial values for the weights and adjusted dependent variable can be calculated. By performing weighted least squares regression on the adjusted dependent variable, estimates of $\hat{\beta}$ are obtained which lead to revised estimates of $\hat{\eta}$ and $\hat{\pi}$, the weights and adjusted dependent variable. The deviance, of which its details are given by (McCullagh and Nelder, 1989), is used to decide whether iteration should stop.

Appendix C

Some Issues Regarding Causality

*****

C1 Assignment Mechanism as Locally Independent

An assignment mechanism is locally independent if (Rubin, 1991; Mattei, 2004):

$$\begin{aligned}Pr(\mathbf{Z}|\mathbf{X},\mathbf{Y}(C),\mathbf{Y}(T) = g(\mathbf{Z},\mathbf{X})\prod_{i=1}^{N}&(Pr(Z_i = T|\mathbf{X},\mathbf{Y}(C)),\mathbf{Y}(T))^{I[Z_i=T]}\\ &\times (1 - Pr(Z_i = T|\mathbf{X},\mathbf{Y}(C)),\mathbf{Y}(T))^{(1-I[Z_i=T])}\end{aligned}$$

and

$$Pr(\mathbf{Z}|\mathbf{X},\mathbf{Y}(C),\mathbf{Y}(T)) = Pr(Z_i = T|\mathbf{X}_i, Y_i(C), Y_i(T)) \qquad \text{(C1.1)}$$

for all i; where $g(Z, X)$ must be such that

$$\sum_{\mathbf{Z}} Pr(\mathbf{Z}|\mathbf{X},\mathbf{Y}(C),\mathbf{Y}(T)) = 1. \qquad \text{(C1.2)}$$

An example of such assignment mechanisms is a completely randomised experiment (CRD) where M out of N units ($M<N$) are randomly chosen to receive the treatment. Rubin (1991, Page 1220) gives an example of an application of assignment mechanism or a simplified assignment mechanism. For this assignment mechanism

$$Pr(\mathbf{Z}|\mathbf{X},\mathbf{Y}(C),\mathbf{Y}(T)) = \begin{cases} \left(\frac{M}{N}\right)^{I\{Z_1=T\}} \left(\frac{N-M}{N}\right)^{I\{Z_1=C\}} & , \sum I\{Z_i = T\} = M \\ 0 & , Otherwise. \end{cases} \qquad \text{(C1.3)}$$

Most often $M = N/2$, so that half the units receive the active treatment and half receive the control treatment.

Being ignorable and locally independent, a classical randomised experiment is also unconfounded, that is, it does not depend on the potential outcomes (Rubin, 1991; Mattei, 2004; Stuart, 2010):

$$Pr(\mathbf{Z}|\mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)) = Pr(\mathbf{Z}|\mathbf{X}) \tag{C1.4}$$

With an unconfounded assignment mechanism, at each set of values of X_i that has a distinct probability of $Z_i = T$, there is effectively a randomised experiment.

C2 Unbiased Estimator of the Average Difference between Study Arms

Mattei (2004, Page 10) provided the following argument to prove that $\hat{\tau}$ is an unbiased estimator for the typical causal effect τ over the randomisation set .

“Let $\hat{\tau}$ be the observed average difference between the treatment and control group:

$$\hat{\tau} = \bar{Y}_T^{obs} - \bar{Y}_C^{obs} = \frac{\sum_{i=1}^N I\{Z_i = T\} Y_i^{obs}}{\sum_{i=1}^N I\{Z_i = T\}} - \frac{\sum_{i=1}^N I\{Z_i = C\} Y_i^{obs}}{\sum_{i=1}^N I\{Z_i = C\}} \tag{C2.1}$$

This is an unbiased estimator for the typical causal effect τ over the randomisation set.

To show this, first we define the randomisation set to be the set of r allocations that were equally likely to be observed given the randomisation plan. For instance, in a completely randomised experiment with $M < N$ units assigned to treatment, the randomisation set is the collection of $r = \binom{N}{M}$ equally likely possible allocations.

For each of the r possible allocations in the randomisation set, there is a corresponding average difference $\hat{r}$ that would be calculated had that allocation been chosen. If the expectation of these r possible differences equals τ, the average difference $\hat{\tau}$ is called unbiased over the randomisation set for estimating τ. We now show that given randomly assigned treatments, the average difference $\hat{\tau}$ is an unbiased estimate of τ, the typical causal effect for the N units.

By definition of random assignment each unit has a known probability of receiving the active treatment, here assumed constant and equal to p. Hence, the contribution of the i^{th} unit ($i = 1, \cdots, N$) to the average difference $\hat{\tau}$ in p of the r allocations in the randomisation set is $Y_i(T)/(Np)$ and in the other $(1 - p)$ is $-Y_i(C)/((1 - p)N)$. The expected contribution of the i^{th} unit to the average difference $\hat{\tau}$ is therefore

$$p\frac{Y_i(T)}{pN} + (1 - p)\frac{-Y_i(C)}{(1 - p)N}. \tag{C2.2}$$

Summing over all N units we have the expectation of the average difference $\hat{\tau}$ over the r allocations in the randomisation set as

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i(T) - Y_i(C)), \tag{C2.3}$$

which is the typical causal effect for the N units in the trial, τ.

The unbiasedness of the $\hat{\tau}$ estimator for τ, that follows from the random assignment of treatments, is a desirable property because it indicates that on average we tend to estimate the correct quantity, however it hardly solves the problem of estimating the typical causal effect. As yet we have no indication whether to believe $\hat{\tau}$ is close to τ or to any ability to adjust for important information we may possess.

Consider now the other formal advantage of randomisation. We show that randomisation provides a mechanism for making probabilistic statements indicating how unusual the observed difference $\hat{\tau}$ would be under specific hypotheses.

Suppose that the researcher hypothesises exactly what the individual causal effects are for each of the N units and these hypothesised values are $\tilde{\tau}_i, i = 1, \cdots, N$. The hypothesised typical causal effect for the N units is thus

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\tau}_i. \tag{C2.4}$$

Having the $\tilde{\tau}_i$ and the observed $Y_i(T), i \in \{i : Z_i = T\}$, and $Y_i(C), i \in \{i : Z_i = C\}$, we can easily calculate hypothesised values, say $\tilde{Y}_i(T)$ and $\tilde{Y}_i(C)$ for all the N units, and using these, we can calculate an hypothesised average difference between the treatment and control group for each of the r allocations of the N units in the randomisation set (Mattei, 2009).

Since the average of the r average differences between the treatment and control group is the hypothesised typical causal effect, $\tilde{\tau}$, and the r allocations are equally likely, we can make the following probabilistic statement:

Under the hypothesis that the causal effects are given by the $\tilde{\tau}_i, i = 1, \cdots, N$, the probability that we would observe an average difference between the treatment and control group that is as far or farther from $\tilde{\tau}$ than the one we have observed is h/r, where h is the number of allocations in the randomisation set that yield average differences between the treatment and control group that are as far or farther from $\tilde{\tau}$ than $\hat{\tau}$ (Rubin, 1974).

If this probability - also called the significance level for the hypothesised $\tilde{\tau}_i$ - is very small, we either must admit that the observed value is unusual in the sense that it is in the tail of the distribution of the equally likely differences, or we must reject the plausibility of the hypothesised $\tilde{\tau}_i$.

The ability to make precise probabilistic statements about the observed $\hat{\tau}_i$ under various hypotheses without additional assumptions is a tremendous benefit of randomisation especially since $\hat{\tau}$ tends to estimate τ. However, one must realise that these simple probabilistic statements refer only to the N study subjects used in the study and do not reflect additional information that we may also have measured.

In order to make an intelligent adjustment for extra information, we cannot be guided only by the concept of unbiasedness over the randomisation set. We need some model for the effect of prior variable in order to use their value in an intelligent manner. The point of this statement is that when trying to estimate the typical causal effect in the N trial experiment, handling additional variables may not be trivial without a well-developed causal model that will properly adjust for those prior variables that causally affect Y and ignore other variables that do not causally affect Y even if they are highly correlated with the observed values of Y. Without such a model, the researcher must be prepared to ignore some variables he or she feels cannot affect Y and use a somewhat arbitrary model to adjust for those variables that he feels are important.

Appendix D

Steps in Implementing Matching Methods

*****

This section provides an overview of approaches on steps to consider in implementing matching methods as given by Stuart and Rubin (2008) and Stuart (2010)

1. **Choosing the covariates** to be used in the matching process

   An underlying assumption when estimating causal effects using nonexperimental data is that treatment assignment is unconfounded (Rosenbaum and Rubin, 1983a) given the covariates used in the matching process. To make this assumption plausible, it is important to include in the matching procedure any covariates that may be related to treatment assignment and the outcome (D'Agostino, 2007); the most important covariates to include are those that are related to treatment assignment because the matching will typically be done for many outcomes. It is important to include a large set of covariates in the matching procedure.

   Another consideration is that the covariates included in the matching must be proper covariates in the sense of not being affected by treatment assignment (Stuart and Rubin, 2008). It is well-known that matching or subclassifying on a variable affected by treatment assignment can lead to substantial bias in the estimated treatment effect (Stuart and Rubin, 2008; Stuart, 2010). All variables should thus be carefully considered as to whether they are proper covariates. This is especially important in fields such as epidemiology and political science, where the treatment assignment date is often somewhat undefined (Reinisch, Sanders, Mortensen and Rubin, 1996). If it is deemed to be critical to control for a variable potentially affected by treatment assignment, it is better to exclude that variable in the matching procedure and include it in the analysis model for the outcome (Reinisch et al., 1996; Stuart and Rubin, 2008) and hope for balance on it, or use principal stratification methods (Frangakis and Rubin, 2002) to deal with it.

2. **Defining a distance measure**, used to assess whether units are similar in terms of their covariate values. We note that the distance measure will not necessarity be a proper *full-rank* distance in the mathematical sense (Stuart and Rubin, 2008). One extreme distance measure is that of exact matching, which groups units only if they have the same values of all the covariates. Because limited sample sizes (and large numbers of covariates) make it very difficult to obtain exact matches, distance measures that are not full rank and that combine distances on individual covariates, such as propensity scores, are commonly used in practice (Stuart and Rubin, 2008).

   Two measures of the distance between units on multiple covariates are the Mahalanobis distance (Morozova, Elizarova, and Pleteneva, 2013), which is full rank, and the propensity score distance, which is not (Stuart and Rubin, 2008). The Mahalanobis distance [see Morozova et al. (2013)], on covariates X between treated and untreated individuals depends on their observable characteristics $X(T)$ and $X(C)$, respectively, and can be expressed as

$$d\Big(X(T), X(C)\Big) = \Big(X(T) - X(C)\Big)^{'} \Sigma^{-1} \Big(X(T) - X(C)\Big) \qquad \text{(D0.1)}$$

   where Σ can be the true or estimated variance-covariance matrix in the treated group, the control group, or a pooled sample; the control group variance-covariance matrix is usually used. The propensity score distance is defined as the absolute difference in (true or estimated) propensity scores between two units (Stuart, 2010).

   Comparison of the performance of matching methods based on Mahalanobis metric matching and propensity score matching was performed by Gu and Rosenbaum (1993) and Rubin and Thomas (2000), and the findings were that the two distance measures perform similarly when there are a relatively small number of covariates, but propensity score matching works better than Mahalanobis metric matching with large numbers of covariates (greater than 5). One reason for this is that the Mahalanobis metric is attempting to obtain balance on all possible interactions of the covariates (which is very difficult in multivariate space), effectively considering all of the interactions as equally important (Stuart and Rubin, 2008). In contrast, propensity score matching allows the exclusion of terms from the propensity score model and thereby the inclusion of only the important terms (e.g., main effects, two-way interactions) on which to obtain balance.

   These distance measures can be combined or used in conjunction with exact matching on certain covariates. Combining these distance measures with exact matching on certain covariates sets the distance between two units equal to infinity if the units are not exactly matched on those covariates (Stuart and Rubin, 2008). Some of the common matching methods (Stuart, 2010):

(a) Nearest *Neighbour Matching (NNM)* which generally selects k matched controls for each treated unit (often, $k = 1$). The simplest nearest neighbour matching uses a *greedy* algorithm, which cycles through the treated units one at a time, selecting for each the available control unit with the smallest distance to the treated unit. A more sophisticated algorithm, *optimal* matching, minimises a global measure of balance (Rosenbaum, 2002).

Rosenbaum (2002) argues that the collection of matches found using optimal matching can have substantially better balance than matches found using greedy matching, without much loss in computational speed. Generally, greedy matching performs poorly with respect to average pair differences when there is intense competition for controls and performs well when there is little competition. In practical situations, when assessing the matched groups' covariate balance, Gu and Rosenbaum (1993) find that optimal matching does not in general perform any better than greedy matching in terms of creating groups with good balance but does better at reducing the distance between pairs.

When there are large numbers of control units, it is sometimes possible to get multiple good matches for each treated unit, which can reduce sampling variance in the treatment effect estimates. Although one-to-one matching is the most common, a larger number of matches for each treated unit are often possible (Ho et al., 2011). Unless there are many units with the same covariate values, using multiple controls for each treated unit is expected to increase bias because the second, third, and fourth closest matches are, by definition, further away from the treated unit than is the first closest match, but using multiple matches can decrease sampling variance due to the larger matched sample size (Stuart, 2010; Ho et al., 2011). Of course, in settings where the outcome data have yet to be collected and there are cost constraints, researchers must balance the benefit of obtaining multiple matches for each unit with the increased costs. Examples using more than one control match for each treated unit are found in (Rubin and Thomas, 2000).

Another key issue is whether controls can be used as matches for more than one treated unit - whether the matching should be done “with replacement” or “without replacement.” Matching with replacement can often yield better matches because controls that look similar to many treated units can be used multiple times (Kuehl, 2000; Stuart, 2010). In addition, like optimal matching, when matching with replacement, the order in which the treated units are matched does not matter. However, a drawback of matching with replacement may be that only a few unique control units will be selected as matches; the number of times each control is matched should be monitored and reflected in the estimated precision of estimated causal effects (Stuart, 2010).

(b) Limited Exact Matching

Rosenbaum and Rubin (1985) illustrate the futility in attempting to find matching treated and control units with the same values of all the covariates and thus not being able to find matches for most units. However, it is often desirable (and possible) to obtain exact matches on a few key covariates, such as race or sex. Combining exact matching on key covariates with propensity score matching can lead to large reductions in bias and can result in a design analogous to blocking in a randomized experiment (Stuart, 2010).

(c) Mahalanobis Metric Matching on Key Covariates Within Propensity Score Calipers

Stuart (2010) highlights that caliper matching selects matches within a specified range (caliper c) of a one-dimensional covariate X (which may actually be a combination of multiple covariates, such as the propensity score):

$$\left|X_{tj} - X_{cj}\right| \leq c \tag{D0.2}$$

for all treatment/control matched pairs, indexed by j.

Stuart (2010) indicates that with a normally distributed covariate, a caliper of 0.2 standard deviations can remove 98% of the bias due to that covariate, assuming all treated units are matched. Althauser and Rubin (1970) find that even a looser matching (1.0 standard deviations of X) can still remove approximately 75% of the initial bias due to X. Rosenbaum and Rubin (1985) show that if the caliper matching is done using the propensity score, the bias reduction is obtained on all of the covariates that went into the propensity score. They suggest that a caliper of 0.25 standard deviations of the logit transformation of the propensity score can work well in general.

For situations where there are some key continuous covariates on which particularly close matches are desired, Mahalanobis matching on the key covariates can be combined with propensity score matching, resulting in particularly good balance (Rosenbaum and Rubin, 1985; Rubin and Thomas, 2000). The Mahalanobis distance is usually calculated on covariates that are believed to be particularly predictive of the outcome of interest or of treatment assignment (Stuart, 2010).

(d) Subclassification

Rosenbaum (1984) discusses reducing bias due to covariate imbalance in observational studies through subclassification on estimated propensity scores, which forms groups of units with similar propensity scores and thus similar covariate distributions. For example, subclasses may be defined by splitting the treated and control groups at the quintiles of the propensity score in the treated group, leading to five subclasses with approximately the same number of treated units in each.

According to Stuart and Rubin (2008) that work builds on the work by Cochran (1968) on subclassification using a single covariate; when the conditional expectation of the outcome variable is a monotone function of the propensity score, creating just five propensity score subclasses removes at least 90% of the bias in the estimated treatment effect due to each of the observed covariates. Thus, five subclasses are often used, although with large sample sizes more subclasses are often desirable. This method is clearly related to making an ordinal version of a continuous underlying covariate (Stuart, 2010).

Subclassification on the propensity score without subsequent within-strata model adjustment can lead to biased answers due to residual imbalance within the strata (Lunceford and Davidian, 2004).

(e) Full Matching (FM) is an extension of subclassification (Small, Gastwirth, Krieger and Rosenbaum, 2009). It is in FM where the matched sample is composed of matched sets, where each matched set contains either one treated unit and one or more controls, or one control unit and one or more treated units (Stuart and Rubin, 2008; Small et al., 2009). Full matching is optimal in terms of minimising a weighted average of the distances between each treated subject and each control subject within each matched set, or stratum (Austin, 2011).

Because subclassification and full matching place all available units into one of the subclasses, these methods may have particular appeal for researchers who are reluctant to discard some of the control units. However, these methods are not relevant for situations where the matching is being used to select units for follow-up (Stuart and Rubin, 2008).

(f) Weighting Adjustments

Another method that utilises all units is weighting, where observations are weighted by their inverse propensity score (Lunceford and Davidian, 2004). Weighting can also be thought of as the limit of subclassification as the number of observations and subclasses go to infinity. Weighting methods are based on Horvitz-Thompson estimation (Stuart, 2010), used frequently in sample surveys. A drawback of weighting adjustments is that, as with Horvitz-Thompson estimation, the variance can be very large if the weights are extreme (if the propensity scores are close to 0 or 1). Thus, the subclassification or full matching approaches, which also utilise all units, may be more appealing since the resulting weights are less variable (Stuart and Rubin, 2008).

Another type of weighting procedure is that of kernel weighting adjustments, which average over multiple persons in the control group for each treated unit, with weights defined by their distance from the treated unit. Heckman et al. (1998a,b) describe a local linear matching estimator that requires specifying a bandwidth parameter. Generally, larger bandwidths increase bias but reduce variance by putting weight on units that are further away from the treated unit of interest. A complication with these methods is that one needs to define a bandwidth or smoothing parameter that does not generally have an intuitive meaning; Imbens (2004) provides some guidance on that choice (Stuart, 2010). With all of these weighting approaches it is still important to clearly separate the design and analysis stages. The propensity score should be carefully estimated, using either logistic regression, Classification And Regression Tree (CART) discriminant analysis, or generalised boosted models, and the weights set before any use of those weights in models of the outcomes (Stuart and Rubin, 2008; Stuart, 2010).

3. **Diagnosing the matches obtained**

Diagnosing the quality of the matches obtained from a matching method is of primary importance. Extensive diagnostics and propensity score model specification checks are required for each dataset, as discussed by Dehejia

(2005). Matching methods have a variety of simple diagnostic procedures that can be utilised, most based on the idea of assessing balance between the treated and control groups. Although we would ideally compare the multivariate covariate distributions in the two groups that are difficult when there are many covariates, and so generally comparisons are done for each univariate covariate separately, for two-way interactions of covariates, and for the propensity score, as the most important multivariate summary of the covariates.

At a minimum, the balance diagnostics should involve comparing the mean covariate values in the groups, sometimes standardised by the standard deviation in the full sample - ideally other characteristics of the distributions, such as variances, correlations, and interactions between covariates, should also be compared. Common diagnostics include t-tests of the covariates, Kolmogorov-Smirnov tests, and other comparisons of distributions (Austin and Mamdani, 2006). Ho et al. (2007) and Stuart (2010) provide a summary of numerical and graphical summaries of balance, including empirical quantile-quantile plots to examine the empirical distribution of each covariate in the matched samples. Rosenbaum (1984) examines F-ratios from a two-way analysis of variance performed for each covariate, where the factors are treatment/control and propensity score subclasses. Rubin (2001) presents diagnostics that relate to the conditions that indicate when regression analyses are trustworthy. These diagnostics include assessing the standardised difference in means of the propensity scores between the two treatment groups, the ratio of the variances of the propensity scores in the two groups, and for each covariate, the ratio of the variance of the residuals orthogonal to the propensity score in the two groups. The standardised differences in means should generally be less than 0.25 and the variance ratios should be close to one, certainly between 0.5 and 2.

4. **Estimating the effect of the treatment on the outcome**

The analysis of the outcome should proceed only after the observational study design has been set in that the matched samples have been chosen, and it has been determined that the matched samples have adequate balance (Stuart and Rubin, 2008). In keeping with the idea of replicating a randomised experiment, the same methods that would be used in an experiment can be used in the matched data. In particular, matching methods are not designed to "compete" with modeling adjustments such as linear regression, and in fact the two methods have been shown to work best in combination (Stuart and Rubin, 2008). Many authors discuss the benefits of combining matching or propensity score weighting and regression adjustment, for example, Robins and Rotnitzky (1995), Heckman et al. (1997), Rubin and Thomas (2000), (Ichimura and Taber, 2001) and Abadie and Imbens (2006).

The intuition for this is the same as that behind regression adjustment in randomised experiments, where the regression adjustment is used to “clean up” small residual covariate imbalance between the treatment and control groups (Heckman et al., 1997; Ichimura and Taber, 2001). The matching method reduces large covariate bias between the treated and control groups, and the regression is used to adjust for any small residual bias and to increase efficiency (McNamee, 2005). These “bias-corrected” matching methods have been found in Abadie and Imbens (2006) to work well in practice, using simulated and actual data. Ho et al. (2007) show that models based on matched data are much less sensitive and more robust than are models fit in the full datasets.

Some slight adjustments to the analysis methods are required with particular matching methods. With procedures such as full matching, subclassification, or matching with replacement, where there may be different numbers of treated and control units at each value of the covariates, the analysis should incorporate weights to account for these varying distributions (Stuart and Rubin, 2008). When subclassification has been used, estimates should be obtained separately within each subclass and then aggregated across subclasses to obtain an overall effect (Rosenbaum, 1984).

Estimates within each subclass are sometimes calculated using simple differences in means, although empirical (Lunceford and Davidian, 2004) and theoretical (Abadie and Imbens, 2006) work has shown that better results are obtained if regression adjustment is used in conjunction with the subclassification. When aggregating across subclasses, weighting the subclass estimates by the number of treated units in each subclass estimates the average treatment effect for the units in the treated group; weighting by the overall number of units in each subclass estimates the overall average treatment effect for the population of treated and control units.

Causal Model

One of the goals of the social sciences is to understand social phenomena, that is to exhibit the mechanism underlying and bringing them about. This task goes beyond description: to exhibit this mechanism requires identifying causal relations between variables of interest. Causation is one of the most important and contentious issues in social science.

Causal models model the properties of a social system. In particular, they model the relations between the properties or characteristics of the system, which are represented by variables. Social system simply means a given population which is a set of units such as schools, classrooms or social clubs.

In causal modelling, to model the properties of a social system means to give the scheme, or the skeleton, of how these properties relate to each other. In other words, the causal model models the causal mechanism governing the social system. Simply - causal models attempt to explain the variability of the effect variable by means of appropriate covariates.

Index

*****

Bibliography

*****

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235 – 267.

Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., Abad, A., and Renard, D. (2004). Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational Statistics & Data Analysis*, 47:537 – 563.

Adeyemi, T. (2008). Predicting students' performance in senior secondary certificate examinations from performance in junior secondary certificate examinations in ondo state, nigeria. *Middle-East Journal of Scientific Research*, 3(2):73 – 81.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, INC, New York.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, INC, New York, second edition.

Akanbi, M. (2014). Impact of divorce on academic performance of senior secondary students in ilorin metropolis, kwara state. *Research on Humanities and Social Sciences*, 4(19):103 – 107.

Alexopoulos, E. (2010). Introduction to multivariate regression analysis. *HIPPOKRATIA*, 14(Suppl 1):23 – 28.

Althauser, R. and Rubin, D. (1970). The computerized construction of a matched sample. *American Journal of Sociology*, 76:325 – 346.

Altman, D. (1991). *Practical Statistics for Medical Research*. Chapman & Hall/CRC.

Angrist, J. (1990). Lifetime earnings and the vietnam era draft lottery: evidence form social security administrative records. *American Economic Review*, 80:318–335.

Angrist, J. and Krueger, A. (1991). Does compulsory school attendance affect schooling and earnings. *Quarterly Journal of Economics*, 106:979 – 1014.

Aronow, P. and Middleton, J. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1):135 – 154.

Austin, P. (2008). Assessing balancing in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiology and Drug Safety*, 17:1218 – 1225.

Austin, P. (2009). Assessing balancing in measured baseline covariates when using many-to-one matching on the propensity-score. *Statistics in Medicine*, 28:3083 – 3104.

Austin, P. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424. PMID: 21818162.

Austin, P., Grootendorst, P., and Anderson, G. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A monte carlo study. *Statistics in Medicine*, 26:734 – 753.

Austin, P. and Mamdani, M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-ami statin use. *Statistics in Medicine*, 25:2084 – 2106.

Awodun, A., Oni, S., and Oyeniyi, J. (2015). Influence of teacher's gender on secondary school students' performance in physics in ekiti state, nigeria. *International Journal of Innovative Research and Development*, 4(1):72 – 77.

Azzimonti, L., Ieva, F., and Paganoni, A. M. (2013). Nonlinear nonparametric mixed-effects models for unsupervised classification. *Computational Statistics*, 28(4):1549–1570.

Balfanz, R. and Byrnes, V. (2012). Chronic absenteeism: Summarizing what we know from nationally available data.

Bansilal, S., James, A., and Naidoo, M. (2010). Whose voice matters? Learners. *South African Journal of Education*, 30:153 – 165.

Barnard, J. (1999). Developing an indicator system for schools of choice: A balanced scorecard approach. *Journal of School Choice*, 19(1):28 – 31.

Barnett, D. (1997). The effects of early intervention on maltreating parents and their children. In Guralnick, M., editor, *The effectiveness of early intervention.*, pages 147 – 170. Baltimore, MD: Brookes (ERIC Document No. ED414694).

Baum, C. (2006). *An Introduction to Modern Econometrics Using Stata*. A Stata Press Publication, College Station, Texas.

BEC (2011). Botswana General Certificate of Secondary Education: 2011 Examination Summary Results. Technical report, Botswana Examinations Council.

Bergsma, W., Croon, M., and Hagenaars, J. (2009). *Marginal Models: For Dependent, Clustered, and Longitudinal Categorical Data*. Springer, New York.

Briggs, D. (2003). Environmental pollution and the global burden of disease. 68(1):1–24.

Britton, A.and McPherson, K., McKee, M., Sanderson, C., Black, N., and Bain, C. (1998). Choosing between randomised and non-randomised studies: a systematic review. *Health Technology Assessment*, 2(13):1 – 124.

Brown, R., Wohlstetter, P., and Liu, S. (2008). Developing an indicator system for schools of choice: A balanced scorecard approach. *Journal of School Choice*, 2(4):392 – 414.

Burton, P., Gurin, L., and Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine*, 17:1261 – 1291.

Caliendo, M. (2006). *Microeconomic Evaluation of Labour Market Policies*. Springer, Berlin, New York.

Carr, M., McGee, C., Jones, A., Mckinley, E., Bell, B., Barr, B., and Simpson, T. (2004). *The effects of curricula and assessment on pedagogical approaches and on educational outcomes. Prepared for the New Zealand Ministry of Education*.

Carter, R. (2006). Solutions for missing data in structural equation modeling. *Research and Practice in Assessment*, 1(1):1 – 6.

Castaldi, B. (1982). *Educational Facilities: Planning, Modernization, and Management*. Boston: Allyn & Bacon, Boston, 2nd edition.

Chow, S. and Liu, J. (2000). *Design and Analysis of Bioavilability and Bioequivalence Studies. Revised and Expanded*. New York: Wiley, Marcel Dekker, Inc. New York, second edition.

Christoffel, K. K., Scheidt, P. C., Agran, P. F., Kraus, J. F., McLoughlin, E., and Paulson, J. A. (1992). Standard definitions for childhood injury research: Excerpts of a conference report. *Pediatrics*, 89(06):1027 – 1034.

Clark, R. M. (1993). Homework-focused parenting practices that positively affect student achievement. In Chavkin, N., editor, *Families and schools in a pluralistic society*, pages 85 – 105. Albany, NY: State University of New York.

Cleaver, H., Unell, I., and Aldgate, J. (2011). *Children's Needs – Parenting Capacity. Child abuse: Parental mental illness, learning disability, substance misuse, and domestic violence*. TSO (The Stationery office), TSO: London, second edition.

Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24:295 – 313.

Concato, J., Shah, N., and Horwitz, R. (2000). Randomized, controlled trials,observational studies and the hierarchy of research designs. *The New England Journal of Medicine*, 342(25):1887 – 1892.

Corby, B. (2006). *Child abuse: Towards a knowledge base*. England: Open University Press, Berkshire, 3rd edition.

Cozby, P. C. (2009). *Methods of Behavioral Research*. McGraw-Hill, New York, NY, tenth edition.

Creswell, J. (2003). *Research design: Qualitative, Quantitative and Mixed Methods Approaches Leadership and Management*. SAGE, Thousand Oaks.

D'Agostino, R. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statist. Med.*, 17:2265 – 2281.

D'Agostino, R. (2007). Estimating treatment effects using observational data. *JAMA*, 297(3):314 – 316.

De Fraja, G., Oliveira, T., and Zanchi, L. (2010). Must try harder: Evaluating the role of effort in educational attainment. *Review of Economics and Statistics*, 3(92):577.

Dehejia, R. (2005). Practical propensity score matching: A reply to smith and todd. *Journal of Econometrics*, 125:355 – 364.

Dehejia, R. and Wahba, S. (2002). The propensity score-matching methods for non-experimental causal studies. *The Review of Economics and Statistics*, 84(1):151 – 161.

Demirbas, O. and Demirkan, H. (2007). Learning styles of design students and the relationship of academic performance and gender in design education. *Learning and Instruction*, 17:345 – 359.

DETE (2013). Performance insights: School attendance.

DFE (2010). Effective Pre-school, Primary and Secondary Education 3-14 Project (EPPSE 3-14) Final Report from the Key Stage 3 Phase: Influences on Students' Development From age 11 - 14. Technical report, UK Government.

Diez Roux, A. (2002). A glossary for multilevel analysis. *J Epidemiol Community Health*, 56:588 – 594.

Donner, A. (1998). Some aspects of the design and analysis of cluster randomization trials. *Appl. Statist.*, 47(1):95 – 113.

Donner, A. and Klar, N. (1994). Cluster randomisation trials in epidemiology: Theory and application. *Journal of Stat Planning and inference*, 42:37 – 56.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold: London, London.

Donner, A., Piaggio, G., and Villar, J. (2003). Meta-analyses of cluster randomization trials: Power considerations. *Evaluation & the Health Professions*, 26(3):340 – 351.

Drake, P. and Heath, L. (2011). *Practitioner research at doctoral level: Developing coherent research methodologies*. Routledge: Oxon.

Eamon, M. K. (2005). Social-demographic, school, neighborhood, and parenting influences on the academic achievement of latino young adolescents. *Journal of Youth and Adolescence*, 34(2):163 – 174.

Fan, X. and Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review*, 13(01):01 – 22.

Farooq, M., Chaudhry, A., Shafiq, M., and Berhanu, G. (2011). Factors affecting students' quality of academic performance: A case of secondary school level. *Journal of Quality and Technology Management*, VII(II):01 – 14.

Fergusson, D. and Boden, J. (2008). Cannabis use and later life outcomes. *Addiction*, 103(6):969 – 976.

Field, A. (2000). *Discovering Statistics using SPSS for Windows*. SAGE Publications, Inc.

Field, A. (2005). *Discovering Statistics using SPSS for Windows*. SAGE Publications, Inc, second edition.

Fives, A., Russell, D., Kearns, N., Lyons, R., Eaton, P., Canavan, J., and O'Brien, A. (2013). The role of random allocation in randomized controlled trials: Distinguishing selection bias from baseline imbalance. *Journal of MultiDisciplinary Evaluation*, 9(20):33 – 42.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. SAGE Publications, Inc.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58:21 – 29.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Gillies, J. and Quijada, J. (2008). Opportunity to Learn: A High Impact Strategy for Improving Educational Outcome in Developing Countries. Technical report, Washington, DC: USAID.

Gnanamoorthy, K. (2014). Contact of socio economic position of the family and their member achievements in education. *Journal of Economics and Finance*, 3(2):31 – 33.

Goldstein, H. (2011). *Multilevel Statistical Models*. John Wiley and Sons, Ltd., Publication, fourth edition.

Gourieroux, C. and Monfort, A. (1981). On the problem of missing data in linear models. *Review of Economic Studies*, XLVIII:579 – 586.

Grant, B. (1998). The impact of a family history of alcoholism on the relation-ship between age at onset of alcohol use and DSM–IV alcohol dependence: Results of the national longitudinal alcohol epidemiologic survey. *Alcohol Health & Research World*, 2:144 – 147.

Green, D. P. and Aronow, P. M. (2011). Analyzing experimental data using regression: When is bias a practical concern? *Available at SSRN 1466886*.

Greenland, S. (2004). Casual inference and observational studies. In Gelman, A. and Meng, X., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An essential journey with Donald Rubin's statistical family*, pages 3–108. John Wiley & Sons, Ltd.

Greenland, S. and Brumback, B. (2002). An overview of relations among causal modelling. *Int J Epidemiology*, 21:1030 – 1037.

Gu, X. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2:405 – 420.

Gunderman, R. and Sistrom, C. (2006). Avoiding errors in reasoning: An introduction to logical fallacies. *AJR*, 187:W469 – W471.

Hailpern, S. and Visintainer, P. (2000). Odds ratio and logistic regression: further examples of their use and interpretatione. *The Stata Journal*, 3(3):213 – 225.

Hair, J., Anderson, R., Tatham, R., and Black, W. (1998). *Multivariate Data Analysis*. Prentice-Hall International, Inc, fifth edition.

Halpern-Felscher, B., Connell, J., Spencer, M., Aber, J., Duncan, G., Clifford, E., Crichlow, W., Usinger, P., Cole, S., Allen, L., and Seidman, E. (1997). Neighborhood and family factors predicting educational risk and attainment in african american and white children and adolescents. In Brooks-Gunn, J., Duncan, G., and Aber, J., editors, *Neighborhood Poverty*. Russell Sage Foundation, New York.

Hardin, J. and Hilbe, J. (2001). *Generalized Linear Models and Extensions*. Stata Press, College Station, Texas.

Hayes, R. and Moulton, L. (2009). *Cluster Randomised Trials*. Chapman and Hall/CRC.

Haynes, R. (2001). Interventions for helping patients to follow prescriptions for medications: Cochrane database of systematic review. *Issue 1*.

Heckman, J., Hidehiko, I., and Petra, T. (1997). Matching as an economemic evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4):605 – 654.

Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. (1998a). Characterizing selection bias using experimental data. *Econometrika*, 66(5):1017 – 1098.

Heckman, J. J., Ichimura, H., and Todd, P. (1998b). Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65:261 – 294.

Hedges, L. and Newell, A. (1999). Changes in the black-white gap in achievement scores. *Bangladesh e-Journal of Sociology*, 72(02):149 – 182.

Henderson, A. and Mapp, K. (2002). A new wave of evidence: The impact of school, family, and community connections on student achievement. *Austin,* *TX: Southwest Educational Development Laboratory. Retrieved 8/11/2014, from www.sedl.org/connections/resources/evidence.pdf.*

Hernan, M. (2004). A definition of causal effect for epidemiological research. *J Epidemiol Community Health*, 58:267 – 271.

Hernan, M., Alonso, A., Logan, R., Grodstein, F., Michels, K., Willett, W., Manson, J., and Robins, J. (2008). Observational studies analyzed like randomized experiments. *Epidemiology*, 19(6):766 – 779.

Herr, K. and Arms, E. (2004). Accountability and single-sex schooling: A collision of reform agendas. *American Educational Research Journal*, 41(3):527 – 555.

Hijazi, S. T. and Naqvi, S. R. (2006). Factors affecting students' performance: A case of private colleges. *Bangladesh e-Journal of Sociology*, 03(01):261 – 294.

Hilbe, J. M. (2009). *Logistic Regression Models*. CRC Press, A Chapman & Hall Book, New York.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278.

Ho, D., Imai, K., King, G., and Stuart, E. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8).

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analyst*, 15(3):261 – 294.

Hochschild, J. L. (2003). Social class in public schools. *Journal of Social Issues*, 59(4):821–840.

Hofler, N. (2005). Causal inference based on counterfactuals. *BMC Medical Research methodology*, 05:28.

Hollard, P. (1986). Statistics and causal inference. *Journal of American Statistical Association*, 81:945 – 970.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley and Sons, INC, New York, second edition.

Houtenville, A. and Conway, K. S. (2008). Parental effort, school resources, and student achievement. *J. Human Resources*, XLIII(2):437 – 453.

Hox, J. (1995). *Applied Multilevel Analysis*. TT-Publikaties, Amsterdam.

Hughes, R. (2008). Divorce and children: An interview with Robert Hughes, Jr., Ph.D. [Retrieved November 1, 2008].

Ichimura, H. and Taber, C. (2001). Propensity-score matching with instrumental variables. *The American Economic Review*, 91(2):119 – 124.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4 – 29.

Imbens, G. W. and Rubin, D. (2000). Unconfounded assignment. *Notes on Unconfoundedness*.

Ioannidis, J., Haidich, A., Pappa, M., Pantazis, N., Kokori, S., Tektonidou, M., Contopoulos-Ioannidis, D., and Lau, J. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *Journal of the American Medical Association*, 286(7):821 – 830.

Isaakidid, P. and Loannidid, J. (2003). Evaluation of cluster randomised controlled trials in sub-saharan africa. *American Journal of Statistics*, 158:921 – 926.

Jabor, K., Machtmes, K., Kungu, K., Buntat, Y., and Nordin, M. (2011). The influence of age and gender on the students' achievement in mathematics. *Journal of Economics Development & Research*, 5(2):Pv2 – 304.

Jensen, E. (2009). *Teaching with Poverty in Mind: What Being Poor does to Kids' Brains and What Schools can Do About It*. ASCD, Alexandria.

Johnson, K. (2000). The peer effect on academic achievement among public elementary school students: A report of the heritage center for data analysis.

Jones, B. and Kenward, M. (1989). *Design and Analysis of Cross-Over Trials*. Chapman & Hall, London.

Jonh, M. (2013). Derivation of sample size formula for cluster randomized trials with binary responses using a general continuity correction factor and identi

cation of optimal settings for small event rates. *Journal of Data Science*, 11:181 – 203.

Jovanovic, J. and King, S. (1998). Boys and girls in the performance-based science classroom: Who's doing the performing? *American Educational Research Journal*, 35(3):477 – 496.

Julious, S. (2010). *Sample sizes for clinical trials.* Chapman & Hall, London.

Karande, S. and Kulkarni, M. (2005). Poor school performance. *Indian J. Pediatr*, pages 691 – 697.

Kassile, T. (2014). Pass rates in primary school leaving examination in Tanzania: Implication for efficient allocation of resources. *South African Journal of Education*, 34(2).

Kilpatrick, J., Swafford, J., and Findell, B. (2001). Adding it up: Helping children learn mathematics. National Academy Press, Washington, DC.

Klein-Geltink, J., Rochon, P., and Dyer, S. (2007). Readers should systematically assess methods used to identify, measure and analyze confounding in observational cohort studies probability. *J Clin Epidemiol*, 60:766 – 772.

Kramer, A. (1995). Health care for elderly persons - myths and realities. *N Engl J Med.*, 332:1027 – 1029.

Kuehl, R. (2000). *Design of experiments: Statistical principles of research design and analysis*. Duxbury Press, Pacific grove, USA, second edition.

Kurdziolek, M. (2011). *Classroom resources and impact on learning*. PhD thesis, The Virginia Polytechnic Institute and State University.

Kyei, K. and Nemaorani, T. (2014). Establishing factors that affect performance of grade 10 students in high school: A case study of Vhembe district in South Africa. *Journal of Emerging Trends in Educational Research and Policy Studies*, 5(7):83 – 87.

Lackney, J. A. (1999). Assessing school facilities for learning/assessing the impact of the physical environment on the educational process: Integrating theoretical issues with practical concerns.

Lacour, M. and Tissington, D. (2011). The effect of poverty on academic achievement. *Educational research and Reviews*, 6(7):522 – 527.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963 – 974.

Langford, I., Bentham, G., and McDonald, A. (1998). Multilevel modeling of geographically aggregated health data: A case study on malignant melanoma mortality and UV exposure in the european community. *Statistics in Medicine*, 17:41 – 57.

Larson, H. (1982). *Introduction to probability theory and statistical inference*. John Wiley & Sons, Inc., New York, third edition.

Lee, W.-C. (2011). Bounding the bias of unmeasured factors with confounding and effect-modifying potentials. *Statistics in Medicine*, 30(9):1007–1017.

Letsoalo, M. (2004). Analysis of Cluster Randomised Trials with Discrete Outcomes. Master's thesis, University of Pretoria, Department of Statistics".

Letsoalo, M. and Lesaoana, M. (2010). Analysis of clustered measurements: A comparison of the performance of foundation year students, 1994 cohort with those of direct students, 1995 cohort, at the University of Limpopo, South Africa. In Reading, C., editor, *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics*. ICOTS.

Letsoalo, M. and Lesaoana, M. (2012). Determination of causal effect in observational studies: Analysis of correlated data with binary end-points. *Journal of Mathematics and System Sciences*, 02:119 – 125.

Levy, S. and Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*. John Wiley & Sons, New York, third edition.

Lewallen, S. and Courtright, P. (1998). Determination of causal effect in observational studies: Analysis of correlated data with binary end-points. *Community Eye Health*, 11.

Lu, M. (2005). Propensity score matching with time-dependent covariates. *Biometrics*, 61:721 – 728.

Luke, D. (2004). *Multilevel Modeling*. SAGE Publications, Thousand Oaks.

Lunceford, J. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23:2937 – 2960.

Manski, C. (1990). Non-parametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319 – 323.

Manski, C., Sandefur, G., McLanahan, S., and Powers, D. (1992). Alternative estimates of the effects of family structure during adolescence on high school graduation. *Journal of the American Statistical Association*, 87:25 – 37.

Mapp, K. (2004). Family engagement. In Schargel, F. and Smink, J., editors, *Helping students graduate: A strategic approach to dropout prevention*, pages 91–113. Larchmont, NY: Eye on Education.

Marjoribanks, K. (1996). Family learning environments and students' outcomes: A review. *Journal of Comparative Family Studies*, 27(2):pp. 373–394.

Mattei, A. (2004). *ESTIMATING CAUSAL EFFECTS IN EXPERIMENTAL AND OBSERVATIONAL STUDIES SUFFERING FROM MISSING DATA*. PhD thesis, UNIVERSIT'A DEGLI STUDI DI FIRENZE, DIPARTIMENTO DI STATISTICA "G. PARENTI".

Mattei, A. (2009). Estimating and using propensity score in presence of missing background data: An application to assess the impact of childbearing on wellbeing. *Stat Methods Appl*, 18:257 – 273.

Mayer, S. (2002). *The Influence of Parental Income on Children's Outcomes*. Knowledge Management Group, Ministry of Social Development, Te Manat$\bar{u}$ Whakahiato Ora.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

McKee, M., Britton, A., Black, N., McPherson, K., Sanderson, C., and Bain, C. (1999). Interpretation of evidence: choosing between randomised and non-randomised studies. *British Medical Journal*, 319:312 – 315.

McMillan, J. and Schumacher, S. (1993). *Research in Education: A Conceptual Introduction*. New York: Harper Collins Publishers, New York.

McNabb, D. (2010). *Research Methods for Political Science: Quantitative and Qualitative Approaches*. New York: Sharpel, New York, second edition.

McNamee, R. (2005). Regression modelling and other methods to control confounding. *Occup Environ Med*, 62:500 – 506.

Mendell, M. J. and Heath, G. A. (2005). Do indoor pollutants and thermal conditions in schools influence student performance? a critical review of the literature. *Indoor air*, 15(01):27 – 52.

MES (2014). Choosing Your Secondary Schools for Admission to Secondary One in 2015. Technical report, Ministry of Education Singapore.

Montgomery, D. and Peck, E. (1982). *Introduction to Linear Regression Analysis*. John Wiley and Sons, INC, New York.

Morgan, K. and Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263 – 1282.

Morgan, S. and Whinship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, Cambridge.

Morozova, M., Elizarova, T., and Pleteneva, T. (2013). Discriminant analysis and mahalanobis distance (nir diffuse reflectance spectra) in the assessment of drug's batch-to-batch dispersion and quality threshold establishment. *European Scientific Journal*, 9:8 – 25.

Murray, D. (1998). *Design and Analysis of Group-Randomized Trials*. Oxford University Press, New York.

Murry, D., Varnell, S. P., and Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *AM J Public Health*, 94.

Mushtaq, I. and Khan, S. (2012). Factors affecting students' academic performance. *Global Journal of Management and Business Research*, 12(09):17 – 22.

Naidu, A., Joubert, R., Mestry, R., Mosoge, J., and Ngcobo, T. (2008). *Education management and leadership: a South African perspective*. Cape Town: Oxford University, Cape Town.

Nathans, L., Oswald, F., and Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, 17(9):1 – 19.

Normand, S., Landrum, M., Guadagnoli, E., Ayanian, J., Ryad, T., Cleary, P., and McNeil, B. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity score. *Journal of Clinical Epidemiology*, 54:387 – 398.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. John Wiley & Sons, INC, Hoboken, New Jersey.

Nuzhat, A., Salem, R. O., Hamdan, N. A., and Ashour, N. (2013). Gender differences in learning styles and academic performance of medical students in Saudi Arabia. *Medical Teacher*, 35(s1):S78 – S82.

Nyagura, L. (1991). A comparative analysis of student achievement by school type in Zimbabwean secondary schools. *Journal of Educational Research*, 3(1):43 – 61.

Okoro, C. C., Ekanem, I. E., and Udoh, N. A. (2012). Teacher gender and the academic performance of children in primary schools in Uyo Metropolis, Akwa Ibom State, Nigeria. *Journal of Educational and Social Research*, 2(1):267 – 273.

Okoro, C. C. and Uwah, C. S. (2013). Teacher gender and attitude of primary school pupils to schooling in Uyo Metropolis, Akwa Ibom State, Nigeria. *Universal Journal of Psychology*, 1(2):53 – 58.

Osborne, J. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, 7(1):1 – 3.

Patterson, S. and Jones, B. (2006). *Bioequivalence and Statistics in Clinical Pharmacology*. Capman & Hall/CRC, London.

Peters, H. and Mullis, N. (1997). The role of family income and sources of income in adolescent achievement. In Duncan, G. and Brooks-Gunn, J., editors, *Consequences of Growing Up Poor*, pages 340 – 381. Russell Sage Foundation Press.

Pocock, S., Assmann, S., Enos, L., and Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21:2917 – 2930.

Pocock, S. and Elbourne, D. (2000). Randomized trials or observational tribulations? *New England Journal of Medicine*, 342(25).

Rabe-Hesketh, S. (2005). *Multilevel and Longitudinal Modeling Using Stata*. StataCorp LP, A Stata Press Publication, College Station, Texas.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2:1 – 21.

Ready, D. (2010). Socioeconomic disadvantage, school attendance, and early cognitive development: The differential effects of school exposure. *Sociology of EducationAmerican*, 83(4):271 – 286.

Reinisch, J., Sanders, S., Mortensen, E., and Rubin, D. (1996). In utero exposure to phenobarbital and intelligence deficits in adult men. *Journal of the American Medical Association*, 274:1518 – 1525.

Robins, J. (2005). Sensitivity analysis in observational studies. In Everitt, B. and Howell, D., editors, *Encyclopedia of Statistics in Behavioral Science*, volume 4, pages 1809–1814. John Wiley and Sons, Ltd.

Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122 – 129.

Rosenbaum, P. (1984). From association to causation in observational studies: the role of tests of strongly ignorable treatment assignmen. *Journal of the American Statistical Association*, 79:41 – 48.

Rosenbaum, P. (2002). *Observational studies*. Springer-Verlag, second edition.

Rosenbaum, P. (2005). Sensitivity analysis in observational studies. *Encyclopedia of Statistics in Behavioral Scienes*, 4:1809 – 1814.

Rosenbaum, P. and Rubin, D. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society*, 45:212 – 218.

Rosenbaum, P. and Rubin, D. (1983b). The central role of the propensity score in observational studies for causal effects. *Biomatrika*, 70:41 – 45.

Rosenbaum, P. and Rubin, D. (1985). Contructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39:1809 – 1814.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Ann. Stat. Assoc.*, 84:1024 – 1032.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society*, 53:597 – 610.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688 – 701.

Rubin, D. (1976). Multivariate matching methods that are equal percent bias reducing: some examples. *Biometrics*, 32:109 – 120.

Rubin, D. (1978). Bayesian inference for causal effect: The role of randomization. *Ann. Stat. Assoc.*, 74:318 – 328.

Rubin, D. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318 – 328.

Rubin, D. (1986). Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961 – 962.

Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores. *Ann Intern Med.*, 127:757 – 763.

Rubin, D. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2:169 – 188.

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandanavian Journal of Statistics*, 31(2):161 – 171.

Rubin, D. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*, 26(1):20 – 36.

Rubin, D. and Thomas, N. (2000). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 95:573 – 585.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral statistics*, 2(1):1–26.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, pages 279 – 292.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, pages 1213–1234.

Rubin, D. B. (2000). Statistical inference for causal effects in epidemiological studies via potential outcomes. *Atti Della XL Riunione Scientifica della Societa Italiana Di Statistica. Roma: Societa Italiana di Statistica*, pages 419–430.

Saghaei, M. (2011). An overview of randomization and minimization programs for randomized clinical trials. *J Med Signals Sens*, 1(1):55 – 61.

Schafer, J. and Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *J Psychological Methods*, 13(4):279 – 313.

Schaffer, M. and Schffer, S. (1997). The impact of divorce on the child in the classroom. *Ten Da'at*, 2(1):24.

Schweinsburg, A., Brown, S., and Tapert, S. (2008). The influence of marijuana use on neurocognitive functioning in adolescents. *Curr Drug Abuse Rev*, 1(1):99 – 111.

Seeger, J. and Walker, A. (2007). Use of propensity score technique to account for exposure-related covariates: An example and lesson. *Pain*, 45(10):S573 – S585.

Shadish, W., Clark, M., and Steiner, P. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334 – 1343.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Berkeley: Houghton Mif-flin.

Sibbald, B. and Roberts, C. (1998). Understanding controlled trials crossover trials. *BMJ*, 316:1719 – 1720.

Silverman, S. and Solmon, M. (1998). The unit of analysis in field research: Issues and approaches to design and data analysisl. *Journal of teaching in physical education*, 17:270 – 284.

Sindrup, S., Andersen, G., Madsen, C., Smith, T., Brosen, K., and Jensen, T. (1999). Tramadol relieves pain and allodynia in polyneuropathy: A randomised, double-blind, controlled trial. *Pain*, 83(01):85 – 90.

Skrondal, A. and Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, 68:267 – 287.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman and Hall/CRC, New York.

Small, D., Gastwirth, J., Krieger, A., and Rosenbaum, P. (2009). Simultaneous sensitivity analysis for observational studies using full matching or matching with multiple controls. *Statistics and its Interface*, 2:203 – 211.

Smith, H. (1999). *Children, feelings and divorce*. London: Free Association Books.

Smith, R. (2011). *Multilevel Modeling of Social Problems: A Causal Perspective*. Springer.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction To Basic And Advanced Multilevel Modeling*. SAGE Publications.

Stols, G. (2013). An investigation into opportunity to learn that is available to grade 12 mathematics learners. *South African Journal of Education*, 33(1):1 – 18.

Stone, W. J. (1986). The carryover effect in presidential elections. *The American Political Science Review*, 80(1):271 – 279.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci*, 52:1 – 21.

Stuart, E. and Rubin, D. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. In Osborne, J., editor, *Best Practices in Quantitative Social Science*, pages 155–176. SAGE Publications.

Sullivan, L., Dukes, K., and Losina, E. (1999). An introduction to hierarchical linear modeling. *Statistics in Medicinei*, 18:855 – 888.

Sunday, O. S. and Zaku, J. A. (2013). Gender analysis of students' entry qualification in english language in colleges of education in kwara state. *International Journal of Secondary Education*, 01(05):23 – 25.

Suresh, K. (2011). Sample size estimation and power analysis for clinical research studies. *J Hum Reprod Sci.*, 4(01):08 – 11.

Suresh, K. and Chandrashekara, S. (2012). Sample size estimation and power analysis for clinical research studies. *J Hum Reprod Sci.*, 5(01):07 – 13.

Taras, M. (2005). Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies*, 53(4):466 – 478.

Terr, L. (1991). Childhood traumas: An outline and overview. *American Journal of Psychiatry*, 148(01):10 – 20.

Trojano, M., Pellegrini, F., Paolicelli, D., Fuiani, A., and Di Renzo, V. (2009). Observational studies: Propensity score analysis of non-randomized data. *The International MS Journal*, 16:90 – 97.

Twisk, J. (2003). *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. Cambridge University Press, Cambridge, UK.

Ukoumunne, O., Gulliford, M., and Chinn, S. (2004). On the distribution of random effects in a population-based multi-stage cluster sample survey. *Journal of Official statistics*, 20(3):481 – 493.

Vaida, F., Meng, X., and Xu, R. (2004). Mixed effects models and the em algorithm. In Gelman, A. and Meng, X., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An essential journey with Donald Rubin's statistical family*, pages 253–329. John wiley & Sons, Ltd.

Van der Berg, S. (2007). Apartheid's enduring legacy: Inequalities in education. *Journal of African Economies*, 16(5):849 – 880.

Van der Berg, S. (2008). How effective are poor schools? poverty and educational outcomes in south africa. *Studies in Educational Evaluation*, 34(3):145 – 154.

Van Der Voort, G. and Wood, L. (2014). Assisting school management teams to construct their school improvement plans: An action learning approach. *South African Journal of Education*, 34(3):1 – 7.

Vondra, J. I., Barnett, D., and Cicchetti, D. (1990). Selfconcept, motivation, and competence among preschoolers from maltreating and comparison families. child abuse and neglect. *ERIC Journal No. EJ424532*, 14(04):525 – 540.

Warner, R. (2013). *Applied Statistics: From Bivariate through Multivariate Tehniques studies*. SAGE Publications, second edition.

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley and Sons, INC, New York, third edition.

Welman, C., Kruger, F., and Mitchell, B. (2008). *Research methodology*. Cape Town: Oxford, third edition.

White, K. (1982). The relationship between socioeconomic status and academic achievement. *Psychological Bulletin*, 91:461 – 481.

White-Paper, D. (1995). White paper on education and training.

Wu, H. and Zhang, J.-T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, 97(459):883–897.

Wu, H. and Zhang, J.-T. (2006). *Nonparametric regression methods for longitudinal data analysis: Mixed-Effect Modeling Approaches*. John Wiley and Sons, INC, New York.

Zhang, D., Lin, X., and Sowers, M. (2007). Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome. *Biometrics*, 63(2):351–362.

Zhou, X.-H., Obuchowski, N. A., and McClish, D. K. (2011). *Statistical methods in Diagnostic Medicine*. John Wiley & Sons, INC, New Jersey, second edition.