Modelling student performance in a tertiary preparatory course

Colin Stuart Carmichael

M.Ed, B.Sc(hons)

Submitted in fulfilment of
Master of Philosophy
at the University of Southern Queensland.

November 2006

Abstract

In this dissertation a review of the literature as it applies to the modelling of educational performance data is undertaken. Statistical linear models, including the novel Beta, Tweedie and Tobit regression models, are then applied to the performance data of students who have undertaken a preparatory mathematics course. These models are then critically reviewed and compared with the commonly used standard linear regression model.

Issues that arise from the application of statistical linear models to educational performance data are then explored. For example, the effects of non-Normality, which characterizes educational performance data, and the presence of large numbers of students who fail to complete the course (a characteristic of this particular context), are examined and reported. Both of these effects can violate the underlying assumptions of the standard linear regression model. Simulation studies are then used to assess the appropriateness of the linear model when it is applied under the condition of non-Normality and the presence of large numbers of missing observations.

Findings from this study indicate that issues relating to model effectiveness are clouded in the educational context by typically large values of the error variance (high noise) and the difficulty in finding suitable performance predictors. Educational models of performance typically lack statistical power, so that in many instances it doesn't matter what model is applied to the data. Nevertheless, the study highlights many reasons why models alternative to the standard linear regression model should be applied to such

data. For example, in situations where the effect is not constant over the entire domain of the explanatory variable, a linear model based upon the beta distribution will be much more appropriate. Similarly, in situations where the performance data contains exact zeros (for example the performance of students who withdraw from the course without providing any measure of achievement) it is more appropriate to use a Tweedie linear model than the standard linear regression model.

Certification of dissertation

I certify that the ideas, experimental work, results, analyses, software and conclusions reported in this dissertation are entirely my own effort, except where otherwise acknowledged. I also certify that the work is original and has not been previously submitted for any other award, except where otherwise acknowledged.

Colin Stuart Carmichael Date

Endorsement

Peter Dunn Date

Janet Taylor Date

Acknowledgements

The completion of this project would not have been possible without the generous support of the following people.

- My supervisors Peter Dunn and Janet Taylor.
- My wife Patricia and son Stephen for their continued support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Rationale

The analysis of student performance data has always been a topic of interest to educators, who have traditionally used such analyses to gauge the effectiveness of their teaching. Performance data have traditionally included student results in various assessment tasks, and on a larger scale the number of courses that the student is able to successfully complete. In an age of economic rationalism the analysis of educational performance data is becoming a tool for the assessment of educational institutions themselves. This is certainly the case in the United Kingdom where 'examination results are included as an indicator of standards and quality' for schools (Shaw et al. [86, p. 64]).

Student performance data are also being used extensively in the Australian higher education context. The Australian Department of Education, Science and Training (DEST) currently assesses the performance of Australian tertiary institutions on the basis of a number of indicators [3, 59], including: the attrition rate (a measure of how many students each institution retains) and the progress rate (a measure of how many courses are successfully completed). Consequently research into these aspects of student

performance is a topic of interest in Australia (see for example the recently released report by the Queensland Studies Authority [82]) and is certainly of interest to university management who obtain Government funding that is based, in part, on the University's ability to meet performance benchmarks.

While student performance data at the institutional level tend to rely on the broader measures of attrition and progress rates, student performance data at the course level are usually based on measures of student academic achievement (usually an aggregate of marks on a number of assessment instruments). In fact there is a plethora of studies in which academics have sought to identify and explain the academic achievement of those students who complete their course. It is also relevant in the current context that such performance data include information relating to those students who do not complete the course. That is, the performance data at the course level should include measures of student academic achievement and measures relating to the actual progression of students. Currently the analysis of student performance data at the course level rarely if ever seeks to address both aspects of academic performance.

It is of particular interest to develop models that can explain the variation that occurs in student performance data. For example, at the institutional level Vincent Tinto's 'Model of Institutional Departure' [96] attempts to explain why students leave their study. At the course level, models are created that may explain the variation in, for example, student grades. Such models are either developed or confirmed using statistical techniques.

The use of statistical models in the educational context, although commonplace, is arguably more problematic than their use in other contexts. One reason for this is that many of the factors that are known to contribute to the variation in educational performance data are difficult to measure. Difficulties also arise as many of the models used for this purpose are based on a number of assumptions, for example those related to the distributional features of the performance measure itself. It is more the norm rather than

the exception that educational performance data do not meet the assumptions of the models that are applied to them. Research conclusions that are based on poor models are likely to be flawed; consequently it is in the best interests of academic rigor to investigate methods for overcoming problems with the data and/or the model.

This project seeks to investigate statistical methods for modelling student performance data (including both academic achievement data and progression data) at the course level. More specifically the project aims to investigate the appropriateness of statistical models that are currently applied to student performance data at the course level. The project also seeks to assess the implications to research in general of ignoring the underlying assumptions that commonly used statistical models in this context are based upon.

1.2 Specific project objectives

1. To conduct an extensive and timely review of the literature as it relates to the statistical modelling of student performance data specifically at the course level.

2. To apply and critically investigate common modelling tools, through their application to the performance data of students undertaking a tertiary preparatory mathematics course.

3. To investigate the appropriateness of standard linear regression techniques in the educational context, especially in situations where model assumptions are violated.

1.3 Dissertation Outline

In Chapter 2 a review of the literature as it relates to the modelling of educational performance in a tertiary context takes place. The nature of

educational performance data is investigated as are problems associated with the measurement of such data. The review then investigates common factors that may be included in models of educational performance data. Following this, some of the statistical techniques currently employed are reviewed and the proposed methodology for the current study is justified. This review also highlights the unique aspects of the context in question.

In Chapter 3 the methodology used in this study is developed. The theory of generalized linear models is reviewed in this chapter as it applies to student performance data. Alternative methods for modelling education performance data are also reviewed. The final part of the chapter then explores simulation methods to assess the appropriateness of statistical models.

Chapter 4 deals with the specific results of this study. In the early sections of this chapter the specific data-set is introduced. A number of regression models are then applied to this data-set, based in part on the particular performance that is being modelled. The final part of this chapter then validates these models on an independent data-set and this in turn serves as a means of evaluating the models.

Chapter 5 deals with the issue of violations in the assumptions underlying the use of the standard linear regression model. The first part of the chapter deals with violations in the assumption that errors in the standard linear model are normally distributed. The second part of the chapter deals with the assumption that observations are selected randomly, and in particular explores the issue of missing data, or the data obtained from students who fail to complete the course.

Chapter 6 provides a summary of this study's results and attempts to address each of the specific objectives outlined in Section 1.2.

Chapter 2

Literature Review

In the previous chapter the rationale and specific objectives for this research were outlined. It was noted that this study specifically seeks to investigate models of student educational performance in a tertiary preparatory setting. Educational performance in this study was defined to include measures of both student progression in a course and their subsequent achievement.

In this chapter a review of the current literature is undertaken as it relates to the modelling of educational performance in general, and in the specific context of tertiary education. In particular, measures of student performance are discussed as are possible predictors of performance. Factors that frequently mediate the relationship between predictor and performance are also examined. This chapter critically reviews the statistical methods currently used to create models of student educational performance. The final section of this chapter then examines the underlying assumptions of these models as they relate specifically to the context of educational performance of students in a tertiary preparatory context.

2.1 Measuring performance

While the response variable in some scientific settings is reasonably easy to measure, this is not the case in the educational context. Measures of educational performance are complex and are subject to, for example, issues relating to test reliability and validity. In this section some of these issues will be presented, not with the intention of providing a solution but merely to heighten an awareness of the problems surrounding the interpretation of models based on educational performance data.

Before discussing issues relating to performance it is necessary to define the type of educational performance that will be dealt with in this study. Educational performance data at the course level may include all measures that in someway indicate how a student has performed in the course. The most common measures relate to scores in examinations and tests (termed 'educational achievement'), however other performance measures at this level may include the length of time that the student has remained in the course, how many assessment items they complete and indeed whether the student completed the course. Achievement data should reflect the degree to which students meet and achieve the learning objectives of the course in question. There are, however, three broad categories of learning objectives (Bloom [10]), namely; cognitive (knowledge), affective (feelings and attitudes) and psychomotor (actual physical skills). In educational settings, student achievement data may include measures in any or all of these domains; however in the tertiary education sector such achievement is usually restricted to measures of cognitive objectives. For this reason achievement data in the current study will deal only with the assessment of such objectives.

The measurement of educational achievement data is problematic. It is measured using a variety of techniques, including project and assignment work; however in the Australian Higher Education sector 'there remains a strong culture of testing and an enduring emphasis on the final examination' (James et al. [45]). So in practice such data are often based on an aggregation

of examination, test and perhaps assignment scores. In the United Kingdom Elton [37, p. 35] argues that the measurement of educational achievement in the tertiary sector is carried out by 'people singularly lacking in any knowledge of assessment and measurement theory'. It is unlikely that academics in the Australian setting have any more qualifications in the area of assessment theory than their United Kingdom counterparts. Moreover academics in the Australian tertiary setting have a significant degree of autonomy regarding the standards of their assessment instruments, so that tests and assignments between similar courses in different universities or even within the same university are seldom if every compared or bench-marked (a process which is commonplace in the Australian secondary school sector). With these factors in mind it is possible, even probable, that the assessment instruments used to measure student achievement of course objectives and the way these are aggregated are far from ideal (see for example Carmichael & St. Hill [21]). Indeed measures of student achievement in a course may in fact reflect the student's ability to undertake examinations rather than their knowledge of the course material. Similarly, an instrument used to assess course objectives with one cohort of students may not reliably measure the same objectives with a second cohort of students. While this study does not intend to examine the validity and reliability of instruments used to assess student achievement these issues do need to be noted. Very few, if any, studies that report models of student achievement ever report on the validity and/or reliability of the measures they use to assess this achievement.

Data that measure student achievement of course, program and degree objectives can take many physical forms. At an institutional level the grade point average (GPA) is by far the most commonly used measure of educational achievement (for example De Berard et al. [28], McKenzie et al. [66] and Zeegers [99]), although various aggregates of grades are also used (see for example Lane & Lane [56] and Middleton [68]). The GPA is derived through the numerical coding of achievement grades and then the aggregation (and

averaging) of these grades. Absolutely no attempt is made at the institutional level to compare similar grades across various subjects, so the GPA is very much dependent on the sample of courses that the student takes. At the course level, student achievement of the objectives is often expressed as a percentage (for example Awang et al. [5] and Blackman [8, 9]) or less commonly as a grade. In the later case some researchers correctly treat the grade as an ordinal variable (for example Considine & Zappala [26] and Fielding et al. [39]). While the use of grades to measure student achievement of course objectives better reflects the uncertainty in the fidelity of the measurement scales used, all too often researchers numerically code these grades and then aggregate them mathematically without any consideration of their ordinal nature (for example the GPA). Even within courses, student achievement in some objectives may be 'marked' using a percentile scale, while on other objectives it may be graded. Discrepancies can occur on the final course performance figure depending on how such grades and marks are combined (see for example McDonald & Taylor [64]). The methods used to measure and record educational achievement data will invariably influence the subsequent statistical methods used to model these data. It is likely that incorrect methods of measuring and recording student educational achievement data will ultimately create inaccurate estimates in any models subsequently employed.

Apart from student achievement in the course, their progression through the course and indeed whether they remain in the course are important examples of student performance. All authors noted that attempt to predict academic achievement either fail to comment on their treatment of students who drop from the course (termed 'incomplete students' in this study), or they restrict the scope of their research to include only students who complete the course (see for example Middleton & Gillies [68], Nguyen et al. [72], Chemers et al. [23] and Smith & Schumacher [87]). Arguably such a treatment leads to the loss of useful information and possibly creates inaccurate model estimates.

The recording of performance data relating to student progression through the course can be done in several ways. For example a dichotomous variable could be used to record whether the student has remained in the course for its duration. Another possible measure is a variable that records the duration of time that the student remains in the course. Alternatively the number of assessment items completed by the student may provide more performance information than merely whether they complete the course or not. Such a measure could be treated as a polytomous variable d, which is a measure of the student's contribution, so that:

$$d = \begin{cases} 0 & \text{if the student submitted no assessment items} \\ 1 & \text{if the student only submitted the first assessment item} \\ \vdots & \\ K & \text{if the student submitted all } K \text{ assessment items} \end{cases}$$

In a sense, these alternative measures of student performance are more objective and measurable than achievement.

In this section problems associated with the measurement and subsequent recording of student performance data at the course level were presented. Issues of test reliability and validity were discussed and the fact that these are generally not addressed in the literature was noted. The various methods used to record and measure student performance data were discussed as were the problems associated with their use. It was noted that performance data at the course level could include both measures of student achievement in the course objectives and their progression through the course itself. Having established the importance of aspects regarding the measurement and recording of educational performance, it is necessary to identify factors that might predict this performance and the possible interactions between these

factors. The following section reviews the literature as it relates to predictors of educational performance.

2.2 Predicting performance

In this section a review of the literature is undertaken in order to identify possible predictors of educational performance. Prior to doing this, it is necessary to reproduce a theoretical framework that attempts to simplify and organize the multitude of constructs associated with the human psyche. (A construct is a theoretical concept that is usually measured through a factor analytic solution of a subject's response to a number of questions/stimuli.)

Psychologists have for centuries believed that the human mind consists of three broad categories; affection, conation and cognition (cited in both Miller [69] and Snow et al. [91]). The affective domain deals with human temperament and emotions, while the cognitive domain, for many years predominant in educational research, deals with knowledge and skills. The conative domain includes the motivation behind a person's actions and their volition (ability to take action). Snow et al. [91, p. 247] propose a taxonomy of individual differences which they suggest is 'a provisional lattice on which to hang theories, hypotheses and findings' (see Figure 2.1). The three domains mentioned above form the basis of this taxonomy. Further, Snow et al. [91] suggest that within each domain there exists a continuum of possible constructs. So that in the affective domain constructs may range from those that measure 'temperament' (a more stable aspect of affection) to those that measure 'emotion'. Similarly in the conative domain, constructs may range from those that measure 'motivation' to those that measure 'volition'. In the cognitive domain knowledge ranges from procedural (skills) to declarative.

While this study seeks to model performance in the cognitive domain, individual differences in such performance have been shown to be dependent on theoretical constructs (discussed below) that span some or all of the

<table>
<tr><th colspan="2">Affection</th><th colspan="2">Conation</th><th colspan="2">Cognition</th></tr>
<tr><td>Temperament</td><td>Emotion</td><td>Motivation</td><td>Volition</td><td>Procedural Knowledge</td><td>Declarative Knowledge</td></tr>
<tr><td>Traits of temperament: Activity level Sociability Emotionality</td><td>Characteristic moods:</td><td>Achievement orientations, eg: goal orientation.</td><td>Action controls, eg motivation control.</td><td>Skills</td><td>Knowledge</td></tr>
<tr><td></td><td></td><td>Self-efficacy</td><td>Meta-cognition</td><td></td><td></td></tr>
<tr><td></td><td></td><td colspan="4">Self regulated learning</td></tr>
<tr><td colspan="6">The Big 5 personality types</td></tr>
<tr><td colspan="2">Agreeableness, extraversion, neuroticism.</td><td colspan="2">Conscientiousness</td><td colspan="2">Intellectual openness</td></tr>
</table>

Figure 2.1: Taxonomy of individual difference constructs [91]

three major domains. In the following review, these theoretical constructs will be classified, where possible, into their dominant domain. In addition to measures associated with the cognitive, affective and conative domains, research literature cites many environmental factors that influence student performance (see for example Bean & Metzner [7]).

Accordingly educational performance may be regarded as having four major categories of explanatory factors:

- cognitive factors: those that primarily span the cognitive domain, such as a student's mathematical ability and their general intelligence;
- affective factors: those that primarily span the affective domain, such as student temperaments and emotional responses;
- conative factors: those that primarily span the conative domain, such as a student's self-efficacy and their persistence; and,
- environmental factors: for example, finances, outside employment, etc.

This categorization of the predictors of academic performance is generally supported in the field. Evidence, cited in Hall & Marchant [41], sug-

gest that predictors of first year academic performance broadly relate to measures of prior academic attainment (cognitive), students' application of their ability and skills (affective/conative) and student allocation of resources (conative/environmental). Similarly in studies of academic retention, Bean's model of non-traditional student attrition [7, p. 491], includes three major categories of variables that map onto the above framework. These four broad categories of factors that predict student performance, viz. cognitive, affective, conative and environmental will be described in more detail in the following sections.

2.2.1 Cognitive predictors of performance

It would seem logical to conclude that cognitive factors would explain cognitive achievement. While this may well be true, identifying and then measuring such factors is not easy. Two such factors that are commonly used are measures of student prior knowledge and intelligence.

Prior academic performance

Prior academic performance is a known predictor of academic performance. Such performance is probably the best way that researchers can measure the current knowledge and skills of students. In a 1987 study of 5000 students across five Australian Universities, Power et al. (cited in Zeegers [99]) found that tertiary entrance score was the best predictor of academic performance in a tertiary context, although Murphy, Papanicolaou & McDowal (1999, cited in McKenzie et al. [66]) subsequently found that this relationship is stronger for students with a high tertiary entrance score. The view that prior academic performance is a key predictor of academic performance is supported by Robbins et al. [84]. In their meta-analysis of 109 studies, Robbins et al. found that a combination of high school grades and standardized test scores accounted for the majority of explained variance in statistical

models of performance and retention. In a further example, Middleton & Gillies [68], in a study of performance in year 11 and 12 Queensland students, found that student performance in a small pretest of 20 arithmetic computations was significantly ($r = 0.31$) correlated with later performance in senior mathematics.

Measures of intelligence

Measures of intelligence have also been shown to predict academic performance. Koke & Vernon [52] found in a study of 150 undergraduate psychology students that measures of intelligence (as obtained from the Wonderlic Personnel Test and also the Sternberg Triarchic Abilities Test) were found to correlate positively with academic performance ($r = 0.41$). Although it would seem logical that intelligence should predict achievement in the cognitive domain, this is not always the case. In a study of 89 students, Dispeth [30] was surprised to find a lack of relationship between general intelligence (measured from three scales) and achievement. It was found that only one measure of intelligence correlated with achievement and then only weakly ($r = 0.23$). Measures of intelligence are fraught with difficulty, as intelligence itself is very difficult to define and measure.

2.2.2 Affective predictors of performance

Temperament includes those dimensions in the affective domain that are biologically based, and consequently more stable. Teglasi [95] cites evidence that the dimensions of temperament resemble some of the commonly used 'Big Five' personality dimensions (see McCrae & Costa [62] for more detail on these personality dimensions). These in turn have been shown to predict academic performance (see for example, Nguyen et al. [72]).

Emotion includes those dimensions in the affective domain that are less stable, for example moods. Research in this area has shown that a stu-

dent's mood during learning will influence the quality of their learning. For example, people recall more material when their recall mood is similar to their mood when learning occurred (Bower [14]). Moreover, a learning environment where positive affect has been created will enhance meaningful cognitive organization and processing (Isen, Daubman & Corgogione (1987) cited in Snow et al. [91]).

2.2.3 Conative predictors of performance

Corno [27, p. 14] argues that a primary issue in education is how individuals move from deliberating about and committing to goals (motivation), to regulation, and finally action (volition). Conative predictors of performance are constructs that measure aspects of motivation and/or volition. While the following discussion examines common constructs from these sub-domains separately, it should be noted that in practice most constructs contain some elements that are motivational and others that are volitional.

Motivational predictors

Although a student's cognitive ability is an important predictor of their ultimate educational performance, their motivation is arguably more important. This view is supported by Robbins et al. [84, p.260] who argue that '... contemporary motivational theories are emerging as strong explanatory models of academic achievement'. Students who do not see any worth in their education are very unlikely to achieve a high performance, unless perhaps they are motivated from external sources, such as their family or peers. Two areas of motivation in the educational context relate to how students' perceive themselves as learners and why they wish to achieve. Both of these areas of motivation have produced constructs that have been shown to predict academic performance, and these will be discussed below.

A person is motivated, in part, by how they view themselves; that is their

self-concept. Self-concept is a very broad construct that attempts to measure a student's knowledge of self and an evaluation of the value that they place on their own abilities (Snow et al. [91]). As it is such a broad construct, self-concept is more commonly measured at more specific levels, such as academic self-concept or even mathematical self-concept. Self-efficacy, which is derived from Bandura's social cognitive theory [6], can be regarded as self-concept at the task level (Snow et al. [91]). More specifically, self-efficacy is defined as a person's confidence in their ability to successfully complete a task. The effects of self-efficacy on learning have been extensively tested (see for example: Awang-Hashim et al. [5]; Pajeres & Miller [78], Pajeres & Graham [77]; Lane & Lane [56]; Stevens et al. [92] and Carmichael & Taylor [22]). In particular, a student's self-efficacy consistently predicts their performance in a variety of contexts (see for example meta-analysis by Multon et al. [70] and Robbins et al. [84]).

A person is also motivated to learn according to the value that they place on the outcomes; that is, how they are oriented towards achievement. Goal orientation theory (Dweck & Leggett [35]), for example, proposes that students usually possess one of two goal orientations. Students who are 'learning orientated' place importance in mastering learning as opposed to students who are 'performance orientated', who place importance in achieving a given performance. Dweck & Leggett [35] have demonstrated an association between the types of goals that a student adopts and their performance. In particular students who are 'learning orientated' are more likely to adopt deep learning strategies, be more persistent and consequently outperform students who are 'performance orientated'. Further, Dweck & Leggett have investigated the mediating effects of student beliefs on student goal orientation (see also Ommundsen [74]). They have demonstrated that students who believe that their ability is fixed are more likely to adopt performance goals, while those who believe their ability can improve are more likely to adopt learning goals.

A person may be negatively oriented towards achievement in that they are motivated through a fear of not achieving. Test anxiety is an example of such a negative achievement orientation. It has been shown to have a negative influence on test performance and often results in maladaptive behaviors (see for example Awang-Hashim et al. [5], Higbee & Thomas [43] and Stodolsky [93]). Anxiety, however, is related to measures of self-efficacy (discussed earlier). Bandura [6, p. 236] cites evidence that 'when anxiety correlates with academic performance, the relation usually disappears or is markedly diminished when the influence of perceived self-efficacy is removed'.

Volitional predictors of performance

While students need to be motivated in order to learn and achieve, they must also have the necessary skills to ensure such learning occurs. These skills might include:

- those necessary to deal with any anxiety that might occur once learning commences (termed 'meta-emotional');
- those skills needed to monitor the effectiveness of learning (termed 'meta-cognitive') and;
- those needed to deal with environmental influences, such as distractions.

Students with strong volitional skills are able to regulate their learning. There is an extensive body of knowledge that currently deals with self-regulated learning. These studies repeatedly demonstrate that students with strong skills in the volitional area out-perform their weaker counterparts (for example: Bouffard et al. [13], Hall & Marchant [41], McKenzie et al. [66], Masui & De Corte [61], Pintrich & De Groot [81], and Zeegers [99]).

2.2.4 Environmental predictors of student performance

There are many factors external to the student that are known to influence their behavior and subsequent academic performance. It should be noted, however, that the influence that such factors have on performance will be mediated by the student's conative skills. For example, two students facing the same environmental influence may react differently according to their level of motivation and or volition.

Environmental factors that are known to influence student performance include their ability to access financial resources, influence from 'significant others', and indeed the mode and quality of teaching. In a study of 466 undergraduates, for example, Cabrera et al. [18] found that encouragement from friends and family ($\beta = 0.217$) and financial attributes ($\beta = 0.054$) both influenced levels of student persistence. Similarly, De Berard et al. [28] found that social support was a significant predictor of academic achievement for 204 college students.

Factors relating to a student's background, such as their ethnic origin and socio-economic status, are also know to affect their performance in education. In a study of retention rates at an institutional level, Allen [2, p. 466] found that students from ethnic minorities were more likely to withdraw from the institution than those from non-minority backgrounds. He reported that 'socioeconomic status tends to favor white males and females'. Considine & Zappala [26], based on their study of 3329 Australian disadvantaged children, argue that the influence of socio-economic status is itself mediated by the educational background of parents.

2.2.5 Mediating factors

In the previous sections four broad groups of predictive factors in educational performance were identified and discussed. It is often the case, however, that there exist factors that mediate the predictive relationship between two

variables. Two common mediating factors identified in the literature are gender and age.

In many studies the relationship between predictive factors and performance is mediated by gender. Males often report a greater level of confidence than females, yet do not necessarily produce a greater level of performance (see for example Awang-Hashim et al. [5]). Nguyen [72] found that gender mediated the influence of personality factors on the academic performance of 360 undergraduate psychology students. Similarly, Bouffard et al. [13] in a study of 702 college students, found that female students were more likely to adopt a learning goal orientation than males, who tended to adopt performance goals and that this tended to confound the overall relationship between goal orientation and academic performance. In a study into mathematics anxiety and its relationship with academic achievement, Zettle & Houghton [100] found that males responding to social stereotyping were more likely to under-report their levels of anxiety, again confounding the established negative relationship between anxiety and academic performance.

Age has been shown to be a mediating factor in the prediction of academic performance. Hall & Marchant [41], for example, found that being over the age of 25 had a positive association with the academic performance of 134 British teacher trainees. In another example, Blackman [8] found that the age of nursing students had a direct negative effect ($\beta = -0.16$) on their achievement in a mental health course. Citing evidence from a number of studies, Petrides et al. [79] suggest that the strength of association between cognitive ability and academic performance declines with age.

2.3 Modelling educational performance

In the last section various known predictors of academic performance were discussed. Many of these were theoretical constructs that related to one or all of the three major domains of the mind. Due to the complexity of the mind,

the identification, measurement and then subsequent modelling of variables associated with the mind will be complex. In this section a review of some of the methods used to create such models is undertaken.

There are many methods that are currently used to model educational performance. The method used, however, is governed by several factors.

1. The intended purpose of the model. If the ultimate purpose of the model is to explain as fully as possible the multitude of inter-relationships that may influence student performance, then by necessity the model used will be complex. If on the other hand, the purpose of the model is to assess the influence of relatively few covariates on student performance, then a simpler model may suffice.

2. The nature of the explanatory variables. Many explanatory variables in educational research are latent, that is they are not directly observable. These variables may include psychological constructs such as personality type, anxiety level and intelligence level. In order to measure these variables researchers need to create psychological scales and these usually comprise a series of questions or statements that elicit some numerical level of agreement from the subject. Such scales are invariably subject to measurement error. A recognition of this error in a model often dictates the type of methodology employed.

3. The nature of the dependent variable(s). The nature of the dependent variable, such as the measurement scale upon which it is measured, will also influence the choice of model. If performance can be regarded as continuous, then perhaps a simple regression model could be employed. If, on the other hand, the performance is measured using grades, then an ordinal model of some description will need to be used.

4. The distribution of the dependent variables. Many statistical models assume Normality of variables, and associated model fit tests are based

on this assumption. If the dependent variable and consequently the errors are not Normally distributed other techniques, for example the use of Normalizing transformations, need to be employed.

5. The structure of the data. It is commonplace to assume that observations are independent. In some cases, the data can be regarded as clustered or hierarchical. A common hierarchical structure occurs when data are collected from students who are assigned to various courses from a number of universities. In such a situation it is probable that there is a certain degree of dependence in the observations obtained from students within the one course, as opposed to observations obtained from students within one university (see for example Snijders & Bosker [90]). The hierarchical (or multilevel) nature of the observations, if indeed it exists, will influence the modelling method used.

In this section two main classes of statistical models employed in educational research are examined, namely regression models and structural equation models. The use of each type of model is dependent on satisfactory answers to the above issues.

2.3.1 Complex models involving latent variables

Research has shown that there are many predictors of educational performance (see discussion in Section 2.2). Often a researcher may wish to construct a model that includes several predictors with the purpose of analyzing both the strength of their predictive power and the inter-relationships between the variables. Zeegers [99] for example constructed a model aimed at explaining academic performance that included as explanatory variables: Self-Efficacy, Test Anxiety, Tertiary Entrance Score, English Language Score, Executive Control, Meta-Cognitive Skills and Approaches to Learning. Such models are very complex and often include latent variables. As mentioned earlier, such variables will include measurement error. In the above exam-

ple, all variables with the exception of Tertiary Entrance Score and English Language Score, were latent variables, measured through student responses to psychological scales.

There are two major methods for dealing with such complex models involving latent variables; Structural Equation Modelling (SEM) and Latent Variable Partial Least Squares Analysis (LVPLSA). While both deal with complex models involving latent variables, they differ in their underlying assumptions and purpose. These differences will be outlined below.

Structural equation models

Structural equation modelling (SEM), also known as covariance structure modelling, is based on the work done by Sewall Wright over 80 years ago that introduced path analysis (see Wolfle [98] for a detailed account). SEM is now supported by several specific software packages, for example LISREL developed by Joreskog in 1973. Structural equation models specifically allow for the error that must occur in the measurement of latent variables. In these models, each latent variable may have associated with it several observable variables that have an independent error.

The estimation of model parameters such as the magnitude of these errors, correlations between variables, and path coefficients, is achieved through the comparison of the sample variance/covariance matrix S with the hypothesized model's variance/covariance matrix Σ. More specifically parameter values are chosen that minimize the 'difference' between S and Σ. The most common method used to estimate such parameters is through maximizing the likelihood function, although a generalized least squares method is also popular. Both of these methods assume that all observations are independent and identically distributed, being drawn from a population that is multi-variate Normal. In practice such an assumption is rarely met (Micceri [67]) and violations of this assumption can lead to problems with the statistical tests associated with model fit (Hu et al. [44], Olsson et al. [73]).

In some cases violations of Normality can be rectified through the use of Normalizing transformations of the data. Such violations of Normality in a multivariate situation, though, are often difficult to isolate and the subsequent interpretation of transformed data is in many cases problematic (Hu et al. [44]). Researchers have developed several techniques in an attempt to overcome the problems associated with non-Normality in the data.

1. Asymptotic distributional free techniques have been developed (Browne [16]) and are available in some commercial software packages (for example LISREL). These techniques rely on large sample sizes in order to produce consistent parameter estimates. Estimation of model parameters is done using a weighted least squares method that relies on weights that address the degree of kurtosis of the sampled data. Recent research by Hu et al. [44], however, has cast doubts on the usefulness of these methods. Using Monte Carlo simulations they found that the asymptotic distributional free techniques performed poorly (based on the number of model rejections expected) even for samples as large as 1000. They also found that under some conditions non-Normal data were in fact better analyzed for model adequacy using Normal methods [44, p. 358].

2. Fitting algorithms have been extended to allow for families of distributions that include the Normal distribution. For example the use of multivariate elliptical theory and heterogenous kurtosis theory developed by Kano et al. [49].

3. Bootstrap techniques have been employed with some degree of success to estimate the standard errors of parameters (Nevitt & Hancock, [71]).

Despite the above attempts to address violations to the Normality assumption, Micceri [67, p. 161] asserts that 'extremes of asymmetry and lumpiness are more the rule than the exception' in measures of ability and psychome-

try. The evolution of a partial least squares approach to structural models (discussed in the next section) has attempted to overcome this problem.

Latent Variable Partial Least Squares Analysis

Latent Variable Partial Least Squares Analysis (LVPLSA) is similar to structural equation modelling, in that it deals with latent explanatory variables and caters for extremely large and complex models. The difference, however, is in the treatment of the errors associated with these variables. LVPLSA estimates each latent variable in the model as a composite of its observable variables. Parameter estimations are performed iteratively and unlike SEM which seeks to minimize the distance between S and Σ, LVPLSA minimizes the error variances within each path of the model. Herman Wold (cited in Fornell & Cha [40, p. 74]) who developed this technique, argued that LVPLSA was 'prediction orientated and gave optimal prediction accuracy'. LVPLSA does not assume any distributional properties of observations and therefore is limited in that models cannot be 'tested'. Standard errors of estimates, however, have been calculated using a jackknife method (see Fornell & Cha [40]).

Despite Blackman's assertion that LVPLSA is 'the modelling method of choice' [8] the parameter estimates obtained from this method are 'less sharp' (Joreskog & Wold [50]) and assume both a large number of cases and a large number of indicators for each latent variable. Moreover, McDonald [65] argues that in this method, latent variables are modelled as composites of their observable variables and that in many cases these composites are treated as though they are in fact the latent variables. Because of its lack of reliance on distributional properties this method has become increasingly popular. McDonald [65, p. 240], however, argues that LVPLSA is complex, consists of 'a set of ad-hoc algorithms' and is limited in that model adequacy is unable to be satisfactorily tested.

Joreskog & Wold [50, p. 270] assert that the two approaches to modelling

latent variables described above, should be 'complementary rather than competitive'. Structural equations are ideal when there are strong prior beliefs about the model that need to be tested while LVPLSA is ideal for more exploratory work.

Limitations of complex models

Tests for model fit in structural equation modelling rely on the assumptions that:

- observations comes from a multivariate Normal distribution;
- observations are collected randomly; and,
- observations are independent.

As noted earlier, the non-Normality of educational performance data is more the norm than the exception. Therefore the likelihood of obtaining data from a truly multivariate Normal distribution, as required in SEM, would be quite unlikely. Some research has been undertaken into the consequences of violations in the Normality assumption for SEM (see for example Chou & Bentler [24], Hu et al. [44] and Olsson et al. [73]). This research suggests that such violations have little effect on parameter estimates but significant effects on model fit statistics. As a consequence several alternative model fitting procedures including LVPLSA have been developed in an attempt to circumvent these violations.

Probably less research has been undertaken into violations of independence and randomness. In an educational setting, data are often hierarchical in nature. That is, observations at a student level, for example, may display a certain degree of dependence if all students are in the one course. Similarly observations at the course level may display dependence if all courses are offered at the one University. Ignoring such a hierarchical structure in a model of educational performance is ignoring dependence within the data.

While structural equation models have been modified to cater for hierarchical data (Muthen (1991) cited in Julian [48]) no applied research noted, has considered this potential source for parameter bias.

Obtaining a random sample, irrespective of the method of analysis employed is not easy in many studies. Researchers do not have the luxury of constructing a population frame and then selecting a random sample from this frame. In many cases samples are obtained from a self-select process and no research noted in the area of structural equation modelling has explored the implications of this selection bias on parameter estimates.

2.3.2 Simpler mathematical models

In the last section, techniques for creating complex models of academic performance were discussed. While such models are useful for modelling academic performance, they are often difficult to interpret. Moreover, these models are limited by their inability to meet the underlying assumptions on which they are based. In this section, simpler models of academic performance will be discussed and in particular the work-horse of modelling, simple linear regression.

Many attempts to model the educational performance of students have used simpler mathematical models than those usually accommodated by structural equation models and path models. This is not to say that both of the latter models cannot be used in simpler situations, just that in these simpler situations there are a larger range of possible tools. In simple studies that seek to examine the influence of, say, only one variable on performance, or perhaps the mediating influence of another variable it is not uncommon for researchers to resort to descriptive statistics such as a comparison of means, or a comparison of correlations. Arguably such tools do not constitute a model, but they do provide useful insights into possible causal relations between variables.

Similarly in such situations it is not uncommon for researchers to use

analysis of variance (ANOVA) or analysis of covariance (ANACOVA) methods. Such methods form the basis of simple causal models and have an equivalence in regression models. For this reason, this section will concentrate on regression models that are used or can be used to model educational performance data.

The simplest linear regression model can be expressed:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{2.1}$$

This model used for predicting the variable Y contains a structural component $\beta_0 + \beta_1 x$, and a random component, the residuals ϵ. This model is based upon a number of assumptions:

- the model is correctly specified;
- the expected value of the residuals is zero $E(\epsilon_i) = 0$;
- the variance of the residuals ϵ_i is constant over the range of X, $\text{var}(\epsilon_i) = \sigma^2$, and;
- the observations are independent, $E(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

Apart from these assumptions (known as the Gauss–Markov conditions) it is commonly assumed that the distribution of residuals is Normal (Gaussian) and that the independent variable X is measured without error. Further, when the model is used to make inference about a population, it is assumed that all observations are sampled randomly from the population.

The model shown in Equation (2.1) can be extended to include more predictors and also more response variables. The model can also be modified to cope with violations in the assumptions listed above and many of these have been incorporated into the theory of generalized linear models (discussed in Chapter 3). Some of these modifications will also be discussed in the next sections.

Dealing with measurement error

The simple linear regression model assumes that the explanatory variable X is measured without error. In practice the data used to explain changes in the response variable are often measured with error (referred to by Kruse & Meyer [54, p. 4] as 'vague data'). This is especially the case when one of the variables is a latent variable, which are frequently predictor variables in models of student performance.

It is possible to account for such error, and perhaps the simplest method involves a consideration of the reliability of the 'scale' used to measure the latent variable. In the simple linear regression model shown in Equation (2.1) the estimate of the regression coefficient is given by:

$$\beta_1 = \frac{\text{cov}(X, Y)}{\sigma_X^2} \tag{2.2}$$

where $\text{cov}(X, Y)$ is the covariance of the variables X and Y and σ_X^2 the variance of the variable X. If there is an allowance for error in the variable X, then the variance observed in X can be partitioned to variance attributed to the true measure of X, σ_T^2 and variance attributed to the error in the measurement of X, σ_e^2; that is:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2.$$

This partition of variance means that the estimate for the regression coefficient becomes:

$$\beta_1 = \frac{\text{cov}(X, Y)}{\sigma_T^2 + \sigma_e^2},$$

and the error will cause this estimate to be biased. One way to overcome this bias is to replace σ_X^2 in Equation (2.2) with an estimate for the true variance of the variable, that is $\sigma_T^2 = \hat{\sigma}_X^2 - \sigma_e^2$. The value of $\hat{\sigma}_X^2$ is the observed

variance and the error variance; σ_e^2 can be estimated from considerations of the reliability of the variable X. That is;

$$\sigma_e^2 = \hat{\sigma}_X^2(1 - \rho_{xx})$$

where ρ_{xx} is the reliability coefficient for the measure of the variable X.

Such a method for attenuating the data is of limited use though. Aiken & West [1, p. 150] make the point that when the reliability of the measures are below 0.7 'no correction method can be expected to salvage analyses containing variables so fraught with measurement error'.

Dealing with dependence amongst the observations

In some cases performance data may be collected so that there are some violations of the independence assumption. For example if performance data are collected across the whole university, but done so by course and then by student, there may be some degree of dependence between observations selected from the one course. This may be due to methods of teaching or merely by a similarity of students electing to take that course. When the data are clustered as in this case, analysis of the data must consider this structure. One way to overcome this dependence in the data is to use multi-level regression models. In such an instance, the data is considered at an individual level and then at a course level. In other words:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij}, \tag{2.3}$$

where y_{ij} is the performance of the i^{th} student in the j^{th} course, β_{0j} and β_{1j} the regression coefficients for the j^{th} course and ϵ_{ij} the residual at the individual level.

The simplest multi-level model is one in which only the intercept term varies between groups, the 'random intercept model' (Snijders & Bosker [90]). In a sense this model assumes the effect size β_1 is the same for each course

in the University. In the random intercept model, Equation (2.3) can be modified if it is assumed that the intercept $\beta_{0j} = \gamma_{00} + U_{0j}$, that is, it has a fixed component γ_{00} (the average intercept) and a random component (or group effect), U_{0j}. The model then becomes

$$Y_{ij} = \gamma_{00} + \beta_{1j}x_{ij} + U_{0j} + \epsilon_{ij}. \tag{2.4}$$

Violations of Normality

Estimates of the parameters β_0 and β_1 in the simple linear regression model can be obtained through a process that minimizes the square of the error term ϵ_i. Provided the Gauss–Markov conditions apply, such estimates are known to be the best linear unbiased estimates obtainable (Gauss–Markov Theorem, cited in Sen & Srivastava [85, p. 41]). Tests of hypotheses and confidence intervals for these estimates, however, rely on the Normality assumption. Traditionally violations in this assumption have been addressed through the use of Normalizing transformations of the response variable. For example, it is common to use either a probit or logit transformation for binomially distributed response variables. While such transformations overcome violations in the Normality assumption, they often produce effect estimates that are difficult to interpret. The use of generalized linear models (McCullagh &Nelder [63]) is one way to overcome the problems associated with the use of Normalizing transformations, as these models are based upon a family of distributions that include the Normal distribution. Such models will form the basis of the work in this project and are discussed further in Section 3.1.

Selection bias

As mentioned in Section 2.3.1 it is usually the case that studies dealing with performance at a tertiary level rely on self-selecting surveys. That is, the formal survey with its population frame and randomly selected sample is not usually employed. This is due, in part, to the low numbers that researchers

encounter in the tertiary education context. In a self-selecting survey one can never be sure of the attributes of subjects who have declined to participate. Consequently estimates of the model parameters may be less accurate.

Such inaccuracies in the estimates of model parameters are said to be 'biased'. More formally, a model estimate is biased if there exists a difference between the expected value of the estimate and its true value. For example, suppose the regression coefficients from a large number of similar samples selected from the same population were estimated, and denoted: $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$. The mean of these estimates $\mu_{\hat{\beta}}$ should be equal to the regression coefficient for the complete population of observations β. Poor sampling design can prevent this from occurring and result in biased estimates. With self-selecting surveys, there is the possibility that students who fail to respond to the survey have characteristics in common that may have affected the survey results if they had been included.

While problems with selection bias are difficult to overcome in studies dealing with educational performance, there is one commonly overlooked area that can be addressed. Many studies that seek to model educational achievement at the post compulsory educational level either restrict their study to students who complete the course or simply ignore the data of incomplete students (see for example example Middleton & Gillies [68], Nguyen et al. [72], Chemers et al. [23] and Smith & Schumacher [87]). In the later case the data have been truncated and Long [58] argues that the incorrect use of such data will result in biased estimates in a simple linear model. In fact he demonstrates a simple simulation in which the actual sign of an effect is changed depending on how the omitted data are treated. Methods have been developed in an attempt to overcome the problems associated with truncated data. One such method is the Tobit model (Breen [15]) which is commonly used in the econometric sciences but rarely (if ever) used in educational research.

The Tobit model assumes that there is an underlying latent variable Y^*,

which in an educational context could be a student's propensity to undertake study. It is only until this latent variable exceeds some threshold c that observable data Y can be collected and the student provides some performance data. If unobservable data are assigned the value c when they fall below this threshold (it is left-censored at c) then the model can be expressed:

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > c \\ c & \text{if } y_i^* \leq c \end{cases}$$

Parameter estimates can be calculated using a maximum likelihood approach. In simulated studies (see Breen [15]) these estimates are unbiased.

2.4 The specific study context

In this study, the performance data for a tertiary preparatory course at the University of Southern Queensland (USQ), Australia, is modelled. Such courses are offered to students who cannot gain entry into the university through mainstream channels and form a part of the University's Tertiary Preparatory Program (TPP). The TPP is a major initiative in the University's access and equity program and through this program non-traditional students can gain access to tertiary education. Students can qualify for a 'fee-free' place in the program if they are able to demonstrate, that due to social and cultural circumstances beyond their control, they were unable to fully utilise prior educational opportunities. Consequently the students in the program come from a diverse range of social and cultural backgrounds. Many have had very limited educational experiences. It is often the case that these students have not studied for several years and are unable to copy with the challenges of formal study. In fact the program has an open entry, which means that students do not have to demonstrate requisite academic knowledge and skills. Due to these factors, withdrawal rates are quite high

for courses offered in the program, with in some cases more than 30% of students dropping within the first 4 weeks. Further to this, there are also many students who do not formally drop from a given course yet do not fully participate. These students do not submit all of the assessment items.

Arguably the underlying distribution of achievement scores for students in these courses is far from Normal and problems encountered with non-Normality in broader educational contexts (see Micceri [67]) are exacerbated in this particular context. Figure 2.2 shows the distribution of final total results for the course TPP7181 in the first semester of 2004. This figure confirms the extreme non-Normality of academic achievement data in this context. With such high withdrawal rates in these courses, any attempt to model student achievement without considering data from those students who have withdrawn is very likely to produce biased results.

2.5 Conclusion

In this chapter some of the general issues regarding the modelling of educational performance in a tertiary setting were discussed and in particular an extensive review of current literature as it applies to this subject was presented. In Section 2.1 problems with the actual measurement of educational performance were discussed. These problems related to issues of test reliability and validity, which are so often ignored by the literature.

Section 2.2 of the chapter examined predictors and mediating factors of performance that have commonly been used in the literature. Commonly used predictors were broadly classified as cognitive, affective, conative and environmental, while gender and age were identified as factors that often mediate the relationship between predictors of performance and actual performance. In Section 2.3 of this chapter, statistical tools that are typically employed to model student performance were critically reviewed. It was found that complex models such as structural equation modelling (SEM)

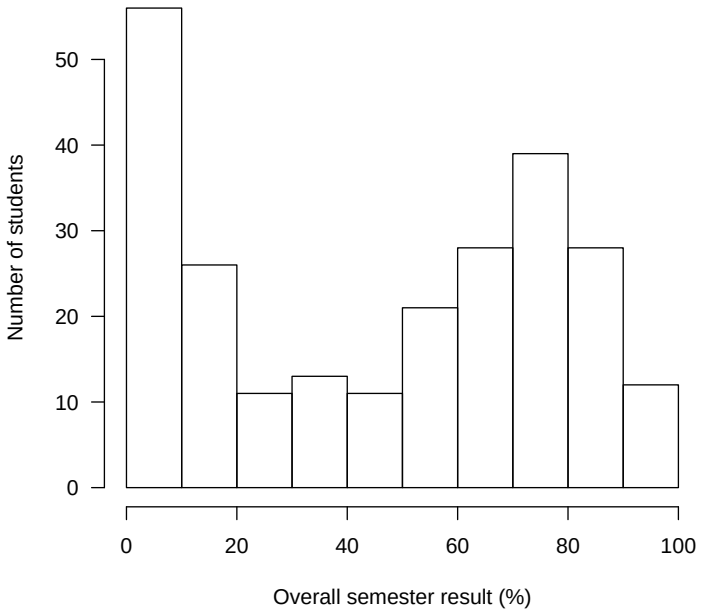

Figure 2.2: Typical achievement distribution

and path modelling (LVPLSA) are very useful if it is the intention to model many predictors and their interactions. Both methods are able to represent complex theoretical models, however they differ in that SEM relies on distributional assumptions which enable researchers to test their model, while LVPLSA does not. While attempts have been made to overcome the problems associated with non-Normality in structural equation models, this still does remain an issue with the use of this statistical technique. In the later parts of this section, simpler statistical tools were discussed. This discussion centred on standard linear regression, which is a popular choice of tool in the modelling of educational performance, probably because it is so well known and relatively easy to use and understand. While regression analysis can

accommodate quite complex theoretical models, it is probably more ideal for those models with few predictors. It relies on distributional assumptions and considerable research has been undertaken in this area in attempts to accommodate non-Normal performance distributions. The issue of selection bias was also discussed in this section and it was pointed out that such bias could be reduced in models of educational performance, if it were possible to retain the results of non-completing students.

In the last section of this chapter the specific context for this study was introduced. In a tertiary preparatory course such as this, a large number of students are likely to withdraw so ignoring their data is likely to create biased results in any study. Similarly, evidence was presented that suggests achievement data from students in such courses are likely to be very non-Normal. Any attempt to model student achievement data in this context will by necessity need to consider both the non-Normality in the data and the data of students who withdraw from the course. As this study seeks to examine only a relatively small number of predictors, it is the intention that the simpler regression models be employed in this study. More specifically the appropriateness of the standard linear regression model will be assessed in situations where there is non-Normality in the data and where there is a large proportion of data from students who withdraw from the course.

Chapter 3

Methodology

In the previous chapter a review of the literature was undertaken as it applied to the modelling of educational performance. It was noted that the application of statistical models in this context is fraught with difficulties, not least being the ability to find suitable explanatory variables. The aims of this study were also developed in the previous chapter.

In this chapter the methodology of this study will be developed. In particular the first section of this chapter will review the theory of generalized linear models. The second section will review techniques for the application of linear models to non-Normal data. The third section of this chapter will then analyze methods for dealing with the exact zeros that occur in educational research. The last section of this chapter will then analyze methods for assessing the appropriateness of a standard linear regression model that is applied to educational performance data that are both non-Normal and that contain exact zeros. As the methodology used in this study is based primarily on the theory of generalized linear models, unless otherwise stated the information discussed in this chapter is obtained from the introductory texts by Dobson [31] and McCullagh & Nelder [63].

3.1 Generalized linear regression models

In Section 2.3.2 simple linear regression models were introduced and the underlying assumptions for these models were presented, one of these being the Normality of the model's random component. It should be noted here that violations in Normality are not an issue if model parameters are fitted using the minimum least squares method (rather than a maximum likelihood approach), however tests for the accuracy of parameter values do rely on Normality. Violations in the Normality of modelled data are not restricted to educational contexts and mathematicians have for years striven to circumvent these through techniques such as the use of Normalizing transformations. The theory of generalized linear models (McCullagh & Nelder [63]) attempts to integrate the many different techniques available for the fitting of linear models to data and will be detailed in this section.

3.1.1 Components of a generalized linear model

A generalized linear model (GLM) has three components:

1. a response variable whose underlying distribution belongs to a family of distributions known as 'Exponential Dispersion Models' (EDMs);

2. a linear combination of covariates used to explain the response $\boldsymbol{\eta}_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, and;

3. a monotonic function (called the link function) linking the mean of the given response's underlying distribution μ_i to the linear predictor $\boldsymbol{\eta}_i$.

The link function incorporates the many types of transformations employed under standard linear regression to Normalize the underlying response distribution. So for example, in modelling a variable that is known to be binomially distributed it is common within the standard linear regression framework to

transform the variable using the logistic transformation (the natural logarithm of the odds ratio). This transformation ensures that the original response defined on the interval $(0, 1)$ is mapped onto the same domain as the linear predictor, in many instances assumed to be $(-\infty, \infty)$. For this reason, in the generalized linear model framework a standard link function for use with binomially distributed data is the logit link, but others also exist.

3.1.2 Exponential Dispersion Models

Exponential dispersion models (Jørgensen [46]) are a family of probability distributions. The distribution of a variable Y is said to belong to this family if its density can be written in the form:

$$f(y; \mu, \phi) = a(y, \phi) \exp\left\{\frac{1}{\phi}[y\theta - \kappa(\theta)]\right\}, \tag{3.1}$$

where $\mu = \kappa'(\theta)$ is the expected value of a random variable Y, $\kappa(\theta)$ is the cumulant generating function for the distribution, and θ a location parameter known as the 'canonical parameter'.

The cumulant generating function, as the name suggests, allows one to generate the cumulants of the distribution, and therefore its mean, variance, skewness (degree of asymmetry) and kurtosis (degree of peakedness). In particular the variance of the random variable Y is given by the expression:

$$\begin{aligned} \text{var}(Y) &= \kappa''(\theta) \\ &= \phi V(\mu) \end{aligned} \tag{3.2}$$

where ϕ is the dispersion parameter of the distribution and $V(\mu)$ the variance function. From Equation (3.2) the variance of an EDM may in fact change as the mean changes. For a Normal distribution, it can easily be shown that $V(\mu) = 1$ so that the variance of the random variable Y is constant (a requirement of the standard linear regression model).

EDMs encompasses a wide range of distributions, including the Normal, binomial, Poisson and gamma distributions. The binomial distribution is used extensively for modelling dichotomously scored performance data (see discussion in Section 3.3.1) while the gamma is used for modelling skewed data (see Section 3.2).

3.1.3 Estimation of parameters

Estimation of the parameters $\boldsymbol{\beta}$ in the linear combination of the covariates can be achieved through the use of a maximum likelihood approach called the method of scoring (see Dobson [31, p.145] for further details). This method is equivalent to the use of an iterative weighted least squares method, where the dependent variable used is a linearized form of the link function obtained through a Taylor series expansion of the link function about $y = \mu$. This is given by:

$$z_i = \eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i}. \tag{3.3}$$

The weights used are the variance of this adjusted dependent variable given by the diagonal matrix W with diagonal elements

$$w_{ii} = \left(\frac{d\mu_i}{d\eta_i}\right)^2 \frac{1}{V(\mu_i)}. \tag{3.4}$$

The equation

$$X^T W X \boldsymbol{\beta} = X^T W \boldsymbol{z} \tag{3.5}$$

is then solved iteratively to produce estimates for $\boldsymbol{\beta}$. The solutions can be obtained without specification of the underlying exponential distribution, provided that the variance function $V(\mu)$ can be specified. This fact becomes important for the fitting of the Tweedie class of distributions discussed in

Section 3.3.3.

Most common software packages, for example SPSS, SPlus and R have in-built routines for fitting generalized linear models. In this project, packages developed within R [83] are used.

3.1.4 Diagnostics associated with GLMs

It is essential that there be some mechanism for establishing the 'goodness of fit' of any model applied to data. In most cases, these diagnostic measures are based in some way on the difference between the value predicted by the model $\hat{y}_i$ and the observed value y_i. This difference $\hat{y}_i - y_i$ is called the raw residual. In simple linear regression models, an assessment of model adequacy is based on the sum of the squares of these residuals. The analogous measure in generalized linear models is called the deviance.

Deviance

The deviance of a generalized linear model is based on the likelihood function for the particular model. The likelihood function, as the name suggests, is a function that evaluates the probability of obtaining the observed data for given values of the model's parameters. The deviance is proportional to the ratio of the likelihood function evaluated for the maximal model (one where the number of fitted parameters equals the number of observations) to the likelihood function evaluated for the model in question. More formally the deviance is defined as

$$D(y, \hat{\mu}) \quad = \quad 2\phi[\ell(y; y) - \ell(\hat{\mu}; y)],$$

where $\ell(y; y)$ denotes the natural logarithm of the likelihood function for the maximal model and $\ell(\hat{\mu}; y)$ the natural logarithm of the likelihood function for the model in question.

The scaled deviance is defined as the deviance divided by the dispersion parameter, that is

$$\begin{aligned} D^*(y,\hat{\mu}) &= \frac{D(y,\hat{\mu})}{\phi} \\ &= 2[\ell(y;y) - \ell(\hat{\mu};y)]. \end{aligned}$$

The scaled deviance for a generalized linear model with N observations and p linear predictors is known to be asymptotically chi-square distributed with $(N-p)$ degrees of freedom (Dobson, [31, p.58]), that is:

$$D^*(y,\hat{\mu}) \sim \chi^2_{N-p}.$$

Consequently the variability in the scaled deviance can be assessed against this chi-square distribution. The scaled deviance, however, is based on knowledge of the dispersion parameter ϕ. If this is unknown, as is usually the case, then an estimate of ϕ such as the mean deviance estimator (see Equation (3.6)) is commonly used. This then means the use of a χ^2_{N-p} distribution to model scaled deviance becomes less accurate and no longer appropriate.

Model fit in the generalized linear model framework is often assessed through the comparison of one model with another, rather than the deviance of the model in question against the relevant chi-square distribution. The difference in deviances scaled by the mean deviance estimator, which is defined

$$\tilde{\phi} = \frac{1}{N-p} D(y,\hat{\mu}), \tag{3.6}$$

can instead be used. In particular, the difference in deviance between two models, one with p parameters and the other with q parameters $(q < p)$ divided by $(p-q)$ and then scaled by the mean deviance estimator (rather than ϕ) is known to be an F-distribution with $(p-q, N-p)$ degrees of

freedom. That is:

$$\frac{\left(D_p(y,\hat{\mu}) - D_q(y,\hat{\mu})\right)/(p-q)}{\tilde{\phi}} \quad \sim \quad F_{(p-q),(n-p)},$$

where $D_p(y,\hat{\mu})$ is the deviance of the model with p parameters and $D_q(y,\hat{\mu})$ the deviance of the model with q variables.

Consequently goodness of fit in generalized linear models can be assessed in much the same way as analysis of variance in standard linear regression. In fact it is common to compare models using this analysis of deviance method.

Residuals

Another common method for analyzing model adequacy is to analyze (often graphically) the residuals associated with each observation. In the case of generalized linear models there are a number of possible residuals that can be used:

1. Deviance residuals are, as the name suggests, based on the deviance and are defined as:

$$r_i^D \quad = \quad \text{sign}(y_i - \hat{\mu_i})\sqrt{d(y_i,\hat{\mu_i})}, \tag{3.7}$$

where $d(y_i,\hat{\mu_i})$ (termed the unit deviance) is the contribution that each observation y_i makes to the overall deviance and the sign function is defined:

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Deviance residuals are approximately Normally distributed with mean 0 and standard deviation ϕ.

2. Pearson residuals are the raw residual scaled by the estimated standard deviation of the response variable. More formally, the Pearson residual is defined as:

$$r_i^P = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)}}, \tag{3.8}$$

where $V(\mu_i)$ is the variance of the hypothesized underlying distribution of y_i (not the sample variance). The distribution of r_i^P is often very skewed for non-Normal data, and so is limited in its usage.

3. Quantile residuals (Dunn & Smyth [34]) are defined for only continuous data and are based on the equivalent standard Normal deviate for each observation. They are therefore exactly Normally distributed with mean 0 and variance 1. More formally, a quantile residual is defined as:

$$r_i^Q = \Psi^{-1}\{F(y_i; \mu_i, \phi)\}, \tag{3.9}$$

where $\Psi(.)$ is the cumulative distribution function for the standard Normal distribution and $F(y_i; \mu_i, \phi)$ the distribution function of the variable in question.

Randomized quantile residuals (Dunn & Smyth [34]) introduce a random component into the calculation of quantile residuals when the response is known to be discrete. This randomization ensures that quantile–quantile plots for such models are approximately linear, rather than the series of steps that would appear if the randomized component was not included. Dunn & Smyth recommend that four replications of this randomization process be undertaken and that features not preserved across each replication are artifacts of the randomization itself.

Assessing the adequacy of the model in question, through the use of residuals, is then based on:

1. A comparison of the sample mean and sample variance of the residuals with the theoretical values; and,

2. The use of quantile–quantile plots to assess the Normality of the residuals. The quantile–quantile (QQ) plot is a graphical technique for determining if two data-sets, in this case the residuals and a variable that is known to be Normally distributed, come from populations with a common distribution. A quantile is the fraction of observations below a certain value, so that a 0.3 quantile is the point on the distribution, below which 0.3 of the observations lie. If the distribution is Normal, then the quantiles of the residuals plotted against a Normal random variable, should form a straight line.

3.1.5 Model fit statistics

In linear regression, the R^2 value is a common statistic for assessing the overall model-fit. The R^2 statistic is equal to the percentage of the variance in the response explained by the model and so gives a fairly good assessment of the effectiveness of the model in question. It should be noted here, that these values are typically quite low in models that attempt to explain educational performance. For example Robbins et al. [84] in a meta-analysis of such studies found that on average such models explained 33% of the variation in performance (that is the average R^2 statistic was 0.33).

Hardin & Hilbe [42] list a number of extensions to the basic notion of the R^2 statistic that can be applied to generalized linear models and some of these will be listed here.

1. Effron's pseudo R-square statistic defined as

$$R^2_E = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}. \tag{3.10}$$

   For Normally distributed data this is identical to the R^2 statistic.

2. McFadden's likelihood-ratio statistic, also called the likelihood-ratio index, which is defined as

$$R^2_M = 1 - \frac{\ell(M_\beta)}{\ell(M_\alpha)}, \tag{3.11}$$

where $\ell(M_\beta)$ is the log likelihood of the full model (or equivalently the residual deviance for the full model) and $\ell(M_\alpha)$ the log likelihood of the same model fitted with only the intercept term (or equivalently the residual deviance for the null model). Again, for Normally distributed data this is identical to the standard R^2 statistic.

3. The adjusted likelihood-ratio index, is defined as:

$$R^2_A = 1 - \frac{\ell(M_\beta) - k}{\ell(M_\alpha)}, \tag{3.12}$$

where $\ell(M_\beta)$ and $\ell(M_\alpha)$ are defined as above and k is the number of parameters in the full model.

In this study, R^2 values quoted will be based on the likelihood based statistics mentioned above. Continuing the notion earlier that deviance is analogous to the sum-of-squares, the use of such a statistic would seem to be the most sensible approach for dealing with the wide range of error distributions allowed for in generalized linear models.

3.2 Modelling non-Normal data

In the previous section aspects relating to the fitting and evaluation of generalized linear models were presented. It was noted that generalized linear models are based on a family of distributions that include many distributions, some of which are not Normal. Certainly the gamma distributions shown in Figure 3.1 are one example. Therefore the application of generalized linear

models that are based on non-Normal distributions is a convenient method for dealing with non-Normal data and will be explored in this section.

3.2.1 The application of GLMs to right skewed data

Gamma regression

The gamma distribution is an EDM and consequently the modelling of observations that are gamma distributed can be undertaken readily using the methods associated with generalized linear models and discussed in the last section. An observation that is gamma distributed has a density given by:

$$f(y, \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \quad y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right), \tag{3.13}$$

where $\Gamma(\alpha)$ is the gamma function, defined by the integral:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} \exp(-x)\, dx. \tag{3.14}$$

The gamma distribution can be quite skewed (see for example the top graph of Figure 3.1) and so is able to cater for observations that are non-Normal, at least those that are right skewed. It is also defined for $Y > 0$ so is appropriate for modelling positive non-zero achievement data.

3.2.2 Ordinal regression models

Violations in the Normality of data are not restricted to the use of data that are distributed asymmetrically, for example the gamma distribution mentioned above. Data that are distributed Normally are assumed to be continuous and defined for all values. It may not be reasonable in this context to assume that the instruments used to measure educational performance have the fidelity required to provide continuous data (see earlier discussion in Section 2.1). Student performance data, and in particular achievement data,

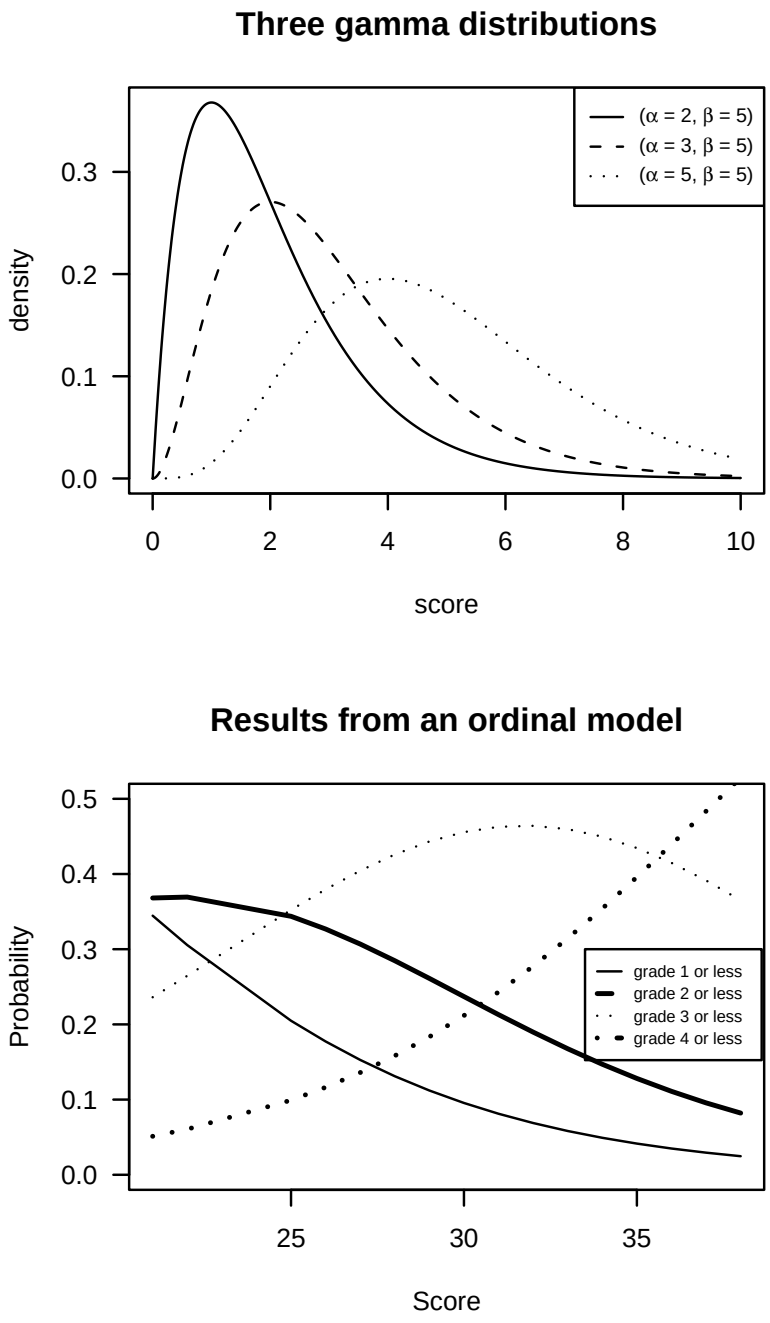

Figure 3.1: Three gamma distributions shown on the top graph and the typical results from an ordinal model on the lower graph

are generally reported in terms of grades or percentages. Both measures are not continuous, with the later being bounded on the interval $(0, 100)$ and the former ordinal. Depending on the number of grades used, the Normal distribution may not be appropriate (nor a good approximation) for modelling the ordinal data produced from grading.

For ordinal achievement data, the existence of an underlying latent performance variable Y^* which is continuous but unobserved may be presumed. The observed ordinal variable Y is defined by:

$$y = \begin{cases} 1 & \text{if } \tau_0 < y^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < y^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < y^* \leq \tau_3 \\ & \vdots \\ k & \text{if } \tau_{k-1} < y^* \leq \tau_k. \end{cases}$$

Assuming that achievement is measured using k grades, then there are $k+1$ boundary points τ_0 to τ_k to be estimated. However it is assumed that $\tau_0 = -\infty$ and $\tau_k = \infty$, so that in effect there are only $k-1$ boundary points to be estimated. It is also assumed that the latent variable Y^* is explained by a linear combination of variables $\boldsymbol{\beta x}_i$ such that $y_i^* = \boldsymbol{\beta}_i \boldsymbol{x}_i + \epsilon_i$. Further it is assumed that the error term ϵ_i is distributed according to a density $f(\epsilon_i)$ with associated cumulative distribution function $F(\epsilon_i)$.

The probability of obtaining an observed value $y_i = k$ is therefore:

$$\begin{aligned} \pi_{ik} &= P(\tau_{k-1} < y^* < \tau_k) \\ &= P(\tau_{k-1} < \beta x + \epsilon < \tau_k) \\ &= P(\tau_{k-1} - \beta x < \epsilon < \tau_k - \beta x) \\ &= F(\tau_k - \beta x) - F(\tau_{k-1} - \beta x) \end{aligned}$$

While it is possible to model the values of π_{ik} directly, 'simple models for the cumulative probabilities are likely to have better properties for ordinal models' (McCullagh & Nelder, p.151, [63]). The cumulative probability that an observed value $y_i \leq k$, is more conveniently given by:

$$\gamma_{ik} = F(\tau_k - \beta x).$$

Further if it is assumed that ϵ has the standard logistic distribution, then the probability of obtaining an observed value $y_i \leq k$ is:

$$\gamma_{ik} = \frac{\exp(\tau_k - \beta x)}{1 + \exp(\tau_k - \beta x)}.$$

This is the basis of the proportional odds model (McCullagh & Nelder [63]) which is an extension of the generalized linear model. Estimation of the parameters τ_i and β_i are obtained using maximum likelihood methods and routines are available in R for achieving this.

Fitting the model to data will result in a plot similar to the lower graph on Figure 3.1 which illustrates the estimated probability curves for data with four ordered categories. It can be seen from this plot that students with a predictor score less than 25 are more likely to achieve a grade of 1 than students who obtain a predictor score greater than 35.

3.2.3 Alternative methods for dealing with non-Normal data

Although generalized linear models can be utilized for modelling data from non-Normal distributions, and this was discussed in the last section, there are alternative linear models in use that can be used for such data. For example, the beta and simplex distributions (discussed below) are not EDMs yet are both suitable for modelling proportions. Student educational achievement data, perhaps not ideally, are often measured on the percentage scale and

can readily be converted to a proportion.

Beta regression

The beta distribution is a two parameter distribution defined on the interval $(0, 1)$ and as such can be used for modelling proportions (see for example Kieschnick and McCullough [51]).

The density of a beta distribution is given by:

$$f(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \tag{3.15}$$

where $\Gamma(.)$ is the gamma function defined earlier in Equation (3.14). The mean and variance of the beta distribution are given by:

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{3.16}$$

and

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{3.17}$$

respectively. The two parameters α and β allow the distribution to take on a number of different shapes, see for example Figure 3.2. Consequently the beta distribution is reasonably versatile, in that it can model both negatively skewed and positively skewed distributions of data. The skewness of the beta distribution is given by:

$$\gamma_1 = \frac{2(\beta - \alpha)}{(\alpha + \beta + 2)} \sqrt{\frac{\alpha + \beta + 1}{\alpha\beta}}. \tag{3.18}$$

From equation (3.18) it can be seen that when the sum of the shape parameters α and β is kept constant and their difference is increased, the skewness should increase also.

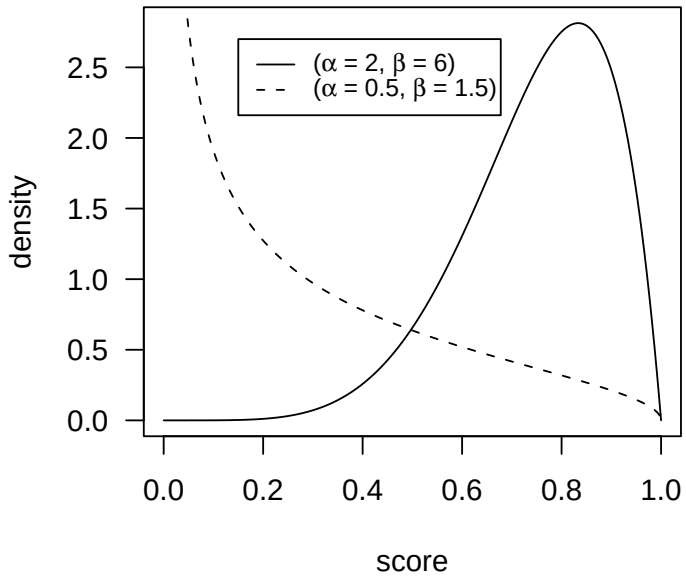

Figure 3.2: Two beta distributions ($\alpha = 6$, $\beta = 2$) and ($\alpha = 0.5$, $\beta = 1.5$)

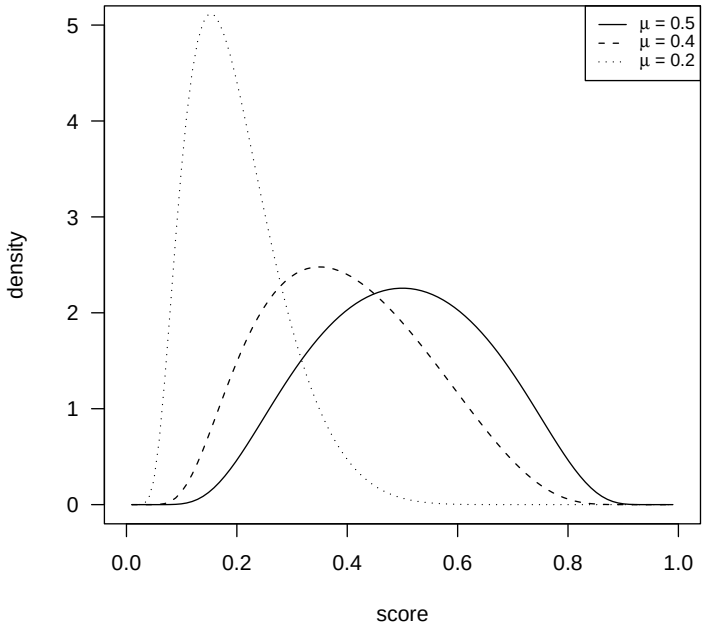

Figure 3.3: Simplex distributions ($\sigma^2 = 2$)

As the beta distribution is not an EDM, linear models based upon this distribution cannot be fitted using the methods associated with generalized linear models. Methods for fitting linear models that are based upon a beta distribution are detailed in Ferrari & Cribari-Neto [38] and closely mirror those used to fit generalized linear models. Software procedures have been developed by Ferrari & Cribari-Neto [29] in R for undertaking this process.

Simplex regression

The simplex distribution (Jørgensen [47]) is similar to the beta distribution in that it is only defined on the interval $(0, 1)$. It also takes on a number of different shapes (see for example Figure 3.3). The density of the simplex can

be written in terms of its unit deviance, consequently analysis of deviance methods can be applied to regression models based upon this distribution.

The density of a univariate simplex distribution is given by:

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y^3(1-y)^3}} \exp\left\{-\frac{1}{2}d(y;\mu)\right\},$$

where the unit deviance is:

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}.$$

Model parameters can be estimated by minimizing the deviance, which is analogous to the least squares parameter estimation method used in simple linear regression.

Numerical methods

In some cases, it is not possible to find a suitable theoretical distribution for the statistical modelling of the variable in question. Parameter estimation can be achieved through a numerical method, known as the bootstrap (Efron & Tibshirani [36]). A bootstrap estimate of the parameters $\boldsymbol{\beta}$ and their standard errors is achieved through the creation of empirical distributions. These distributions are created through sampling of points $(\boldsymbol{x}_i, y_i)$ from the observed distribution of observations. The sampling is done in such a way that each observation has the same chance of being selected. For each empirical distribution so formed, standard regression techniques (such as least squares) are then employed to obtain bootstrap estimates of the regression coefficients $\hat{\boldsymbol{\beta}}^*$. If this procedure is repeated a large number of times, a distribution of $\hat{\boldsymbol{\beta}}^*$ is created and the mean and standard deviation of this bootstrap distribution are excellent estimates for $\boldsymbol{\beta}$ and the standard errors of $\boldsymbol{\beta}$ respectively.

3.3 Modelling exact zeros

In the previous section methods for modelling non-Normal data were discussed. These methods were based primarily on distributions that were non-symmetric, such as the beta, gamma and simplex distributions. These distributions are defined on a domain that does not include zero. In Section 2.4 it was mentioned that a substantial proportion of students within the context of a tertiary preparatory mathematics course withdraw from the course before any measure of performance can be obtained. To ignore the data from students who fail to have any measure of achievement, that is to truncate the data, may be problematic when there is such a large proportion involved. On the other hand to merely code the non-achievement as a zero and include this in a distribution (such as a Normal) that allows only relatively few zeros is problematic. This section details methods for modelling the large proportions of exact zeros encountered in this context. These include the use of logistic regression to model exact zeros, the use of the Tobit model and the use of the Tweedie distribution (Jørgensen [47]).

3.3.1 Logistic regression

The occurrence of exact zeros in an educational context usually occurs when students leave the course without ever having attempted some assessment item. In other words, no measure of achievement is obtained from the student before they exit the course. A standard method for modelling such a situation is to create a dichotomous variable, perhaps called retention, in which students who withdraw from the course are assigned a zero and those that remain assigned a one. Under certain conditions, including the independence of observations, such a variable can be modelled using the binomial distribution. This is an EDM and consequently a generalized linear model can be applied to this situation. It is common to use a logit link (the natural logarithm of the odds ratio) when modelling data that are binomially

distributed, as effects are expressed conveniently in terms of probabilistic statements. Any linear model derived from logistic regression will be of the form:

$$\text{logit}(y_i) \quad = \quad \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where

$$\text{logit}(y_i) \quad = \quad \log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right).$$

3.3.2 The Tobit model

The Tobit model assumes the existence of a continuous underlying latent variable Y^* that only is observed after it reaches some threshold. For example, in a tertiary context, the existence of a continuous latent variable 'propensity to study at a tertiary level' may be hypothesized. It is hypothesized that in this context, students who withdraw from a course without having provided some measure of their achievement, have low, but unobservable scores on this latent variable. When the value of this variable reaches, say 0, then it is possible to record an observable score y_i.

The Tobit model assumes that $y_i^* = \boldsymbol{x_i}^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i$, and that $\epsilon_i \sim N(0, \sigma^2)$. Further, the observable variable y is related to the latent variable according to:

$$y_i = \begin{cases} y_i^* = \boldsymbol{x_i}^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Since ϵ is distributed $N(0, \sigma^2)$ it follows that ϵ/σ is distributed $N(0, 1)$. The probability that an observation will be recorded as zero, given values of the

explanatory variable $\boldsymbol{x_i}$ is therefore:

$$\begin{aligned} P(y_i^* \leq 0 \mid \boldsymbol{x_i}) &= P\left(\frac{\epsilon}{\sigma} \leq \frac{-\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma} \mid \boldsymbol{x_i}\right) \\ &= \Psi\left(\frac{-\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}\right), \end{aligned}$$

where Ψ is the Normal distribution function.

The contribution to the likelihood of a zero observation is:

$$P(y_i = 0 \mid \boldsymbol{x_i}) = \Psi\left(\frac{-\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}\right),$$

while the contribution to the likelihood of a non-zero observation is:

$$P(y_i > 0 \mid \boldsymbol{x_i}) = \frac{1}{\sigma}\psi\left(\frac{y_i - \boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}\right),$$

where ψ is the Normal density function. Summing over all values and taking logs, the resulting log likelihood function is given by:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \boldsymbol{x}) &= \sum_{y_i \geq 0} \log \frac{1}{\sigma}\psi\left(\frac{y_i - \boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}\right) \\ &\quad + \sum_{y_i = 0} \log \Psi\left(\frac{-\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}\right). \end{aligned} \tag{3.19}$$

Maximum likelihood estimates of $\boldsymbol{\beta}$ are obtained by maximizing Equation (3.19) and numerical methods are readily available in the software package R to achieve this. The package Survival ([75]) in R also can be used to deal with left censored data (a term used to describe data in this situation).

The Tobit model has been used extensively in the econometrics field (see for example Burkey & Harris [17]), and also in the health sciences (see Austin et al. [4]). No studies noted have used this model in an educational context.

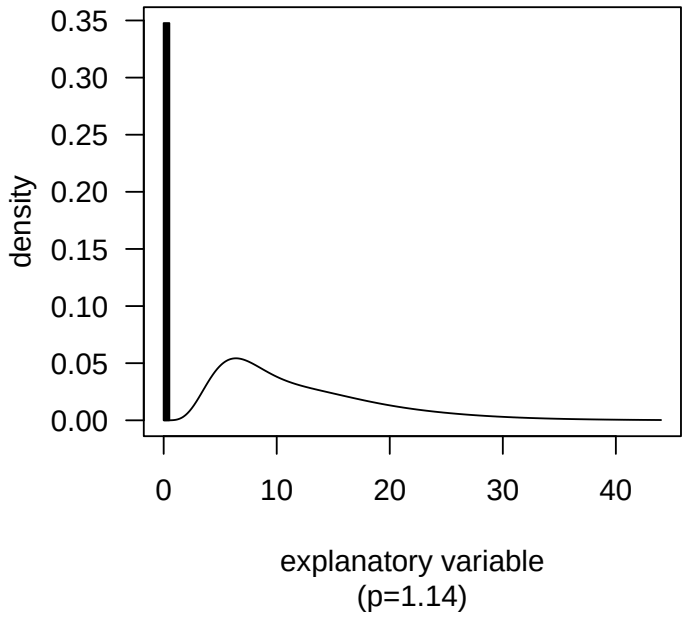

Figure 3.4: A Tweedie distribution ($\mu = 7.8$, $p = 1.14$, $\phi = 6.4$)

3.3.3 Tweedie regression

The Tweedie class of distributions are EDMs with a variance of the form:

$$\text{var}(Y) \quad = \quad \phi\mu^p.$$

This class of distributions encompass quite a number of familiar distributions, including the Normal ($p = 0$), the Poisson ($p = 1$ and $\phi = 1$) and the gamma ($p = 2$). Of particular interest in this study are those with $1 < p < 2$, as they model continuous variables that include exact zeros (see for example Figure 3.4). Jørgensen ([47] p. 140) shows that this sub-group of Tweedie distributions is equivalent to the Poisson sum of a number of

identically distributed gamma random variables. In the educational context, it is hypothesized that there is a number of learning outcomes necessary for successful achievement in the course, with achievement in each being gamma distributed. Theoretically a student's achievement in any one learning outcome has a lower bound of zero and no upper bound (although in practice achievement measures have an upper bound, often 100%). Consequently the achievement distribution for any given learning outcome should be right skewed and possibly gamma distributed. Students' total achievement in the course can then be considered as a Poisson sum of these gamma distributed learning outcomes.

As the Tweedie distributions are exponential dispersion models, methods associated with the fitting of generalized linear models can be applied to this family of distributions. Although Jørgensen ([47] p.141) has specified the form of the density for Tweedie distributions in the range $1 < p < 2$, such a specification is not in closed form and necessitates the use of an infinite series in order to evaluate specific values. Fortunately it is not necessary to specify the density of an EDM in order to estimate the parameters in a generalized linear model (see Section 3.1.3). In such instances the parameter p can be estimated using maximum likelihood techniques developed by Dunn [33] and then standard generalized linear model fit techniques are applied. Software routines for the fitting of Tweedie models have been developed by Smyth [89] and Dunn [32] and are utilized in this study.

3.4 Assessing the appropriateness of the standard linear regression model

In Section 3.2 regression techniques that were based on non-Normal distributions were discussed, in particular the gamma, simplex and beta distributions were found to be possible underlying distributions for student achievement data. In Section 3.3 methods for dealing with exact zeros were also discussed.

These methods included: the hypothesis of the existence of an underlying latent variable and subsequent treatment of zeros as censored; the logistic modelling of these zeros; and the use of a Tweedie distribution to model data that included exact zeros. One of the key objectives of this study, as outlined in Section 1.2, is to assess the appropriateness of the standard linear regression model in the educational context and especially in situations where violations in model assumptions are ignored. In this section methods for assessing this appropriateness will be discussed. These methods rely on the creation of a statistical model that can simulate the types of problems encountered in the actual modelling of educational data.

As mentioned in Section 2.3.1 extremes of asymmetry and lumpiness are the norm in educational achievement data (Micceri [67]). A cursory examination of the achievement data typically generated from a preparatory mathematics course, and shown in Figure 2.2 would support this view. It is hypothesized that such data may be regarded as being generated from more than one underlying student sub-population. In other words, the distribution of results for a particular student population may in fact be a mixture of results from distinct student sub-populations. Such a view is commonly taken by researchers, and the analysis of performance data for sub-populations based on gender or ethnicity are commonplace (see for example Awang-Hashim et al. [5] and Considine & Zappala [26]).

In the current context (see Section 2.4) it is hypothesized that sub-populations be defined on the basis of the extent to which the student completes assessment tasks. The course TPP7181 consists of a number of assessment tasks spaced evenly across the semester. Traditionally students either withdraw without attempting any assessment task, they attempt a sample of the assessment tasks or they may complete all tasks. On this basis there are three major achievement related sub-populations of students:

1. Those students who withdraw from the course without providing any measure of achievement (subsequently referred to as ‘drop-outs’). This

sub-population is quite large in the tertiary preparatory context.

2. Those students who provide some measures of achievement, but who do not complete the entire course (subsequently referred to as 'partials'). These students may complete some assessment items but fail to complete all of the assessment items, and in particular the course examination which carries a substantial proportion of the total course marks.

3. Those students, irrespective of overall achievement, that submit all of the assessment items and in particular the final item (subsequently referred to as 'completes').

Of interest is whether students actually complete the course, so it is often the case that the first two categories above are combined. Such a category of students, that is those who do not complete the course will be referred to as 'incompletes'. The achievement distributions of the three subgroups above are shown in Figure 3.5. In this figure, achievement results for each sub-group in a tertiary preparatory context are shown and labelled, with the achievement distribution for the whole group displayed in the lower right hand graph. It can be seen that the awkward and probably undefinable distribution for the total student group is conceivably a mixture of three smaller and probably definable distributions.

While it is unlikely that any defined distribution could adequately model the distribution of achievement data shown in Figure 2.2 and the bottom right hand corner of Figure 3.5, the distributions discussed earlier in this chapter could conceivably model the distributions of achievement data for the sub-populations described in the previous section and the performance of students generally. More specifically:

1. The overall performance of students, that is whether they drop out or do not drop-out could be modelled using logistic regression (discussed in Section 3.3.1).

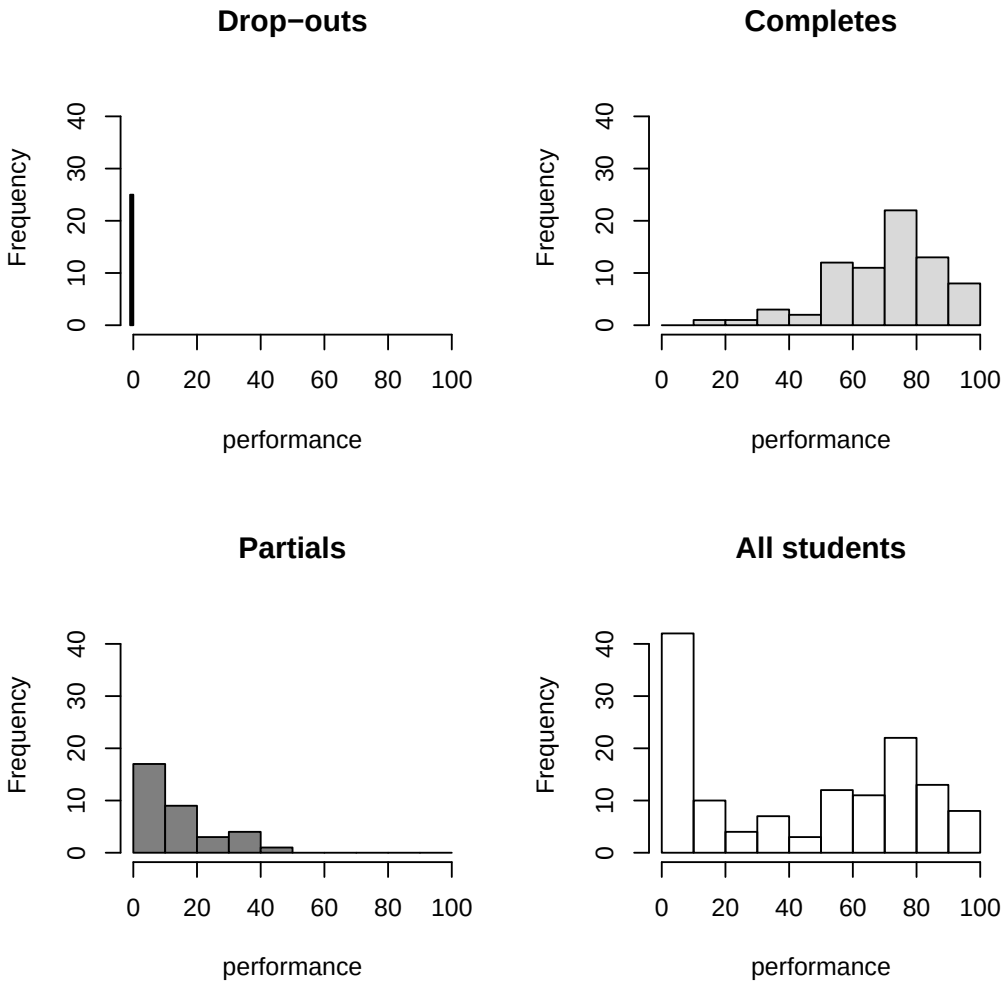

Figure 3.5: Achievement distributions of student sub-groups together with achievement distribution for overall group

2. Achievement data from both students who drop-out and those who partially complete (called collectively 'incompletes') could be modelled using a Tweedie based generalized linear model (see Section 3.3.3) or the Tobit model (see Section 3.3.2).

3. Achievement data from the 'completes' could be modelled using a gamma based generalized linear model (see Section 3.2.1) or regression models based on the beta or simplex distributions (see Section 3.2.3).

Consequently it is possible to generate data from defined distributions that have similar properties to each of the defined sub-population samples. A mixing of these generated data-sets should produce a simulated distribution of achievement data that bears similar properties to that of the sample data. Further, a vector of covariates can be generated that is known to be a linear predictor of this achievement. In such a way, a model can be constructed that attempts to answer some of the major research questions of this study, namely:

1. To what extent do violations in the Normality assumption influence the application of the standard linear regression model to educational data? Moreover, how accurate are the estimated parameters of such a model and how useful is it for prediction purposes?

2. What is the most appropriate way to deal with the data of students who fail to complete the course? Moreover, to what extent does truncation of such data (that is omission of zero results) influence the parameters of linear models that are also based upon Normally distributed data?

Such a model can also consider two other factors that might influence the outcomes to these questions:

1. The influence that the predictive strength of a covariate has on the degree of parameter bias. Typically in a tertiary educational context

linear models only explain approximately 30% of the variation in student performance (Robbins et al. [84]). In other words, given the extent of 'noise' in any educational performance model, do the violations mentioned above really make any difference?

2. Given the existence of the three sub-populations of students, to what extent do variations in the mixing proportions of these sub-populations influence the appropriateness or otherwise of the model in question?

3.5 Conclusion

In this chapter the general methodological framework for this study was developed. In particular generalized linear models were introduced in Section 3.1 and were shown to be an extension of the widely used simple linear regression model that was introduced in Section 2.3.2. Methods used to fit these models were also discussed in this section as were methods used to ascertain model goodness of fit. Generalized linear models accommodate data that are non-Normally distributed, which is one of the basic underlying assumptions of the simple linear regression model. For example data from a right skewed distribution could be modelled within the generalized linear model framework using an underlying gamma distribution (discussed in Section 3.2.1). Generalized linear models can also be applied to data that are not continuous, for example ordinal data (discussed in Section 3.2.2) and data that contain high proportions of zeros (see discussion on Tweedie models in Section 3.3.3).

Whilst generalized linear models provide a versatile framework that can be applied to educational performance data, there are techniques available that may apply to such data but which themselves do not fall within the generalized linear model framework. For example, both the beta and simplex distributions (discussed in Section 3.2.3) are appropriate for modelling data that are defined on the interval (0,1), but neither distribution is an EDM and therefore do not fit within this framework. Similarly, the Tobit

model (discussed in Section 3.3.2) applies to data that may contain exact zeros where it is assumed that censoring has occurred. This model is based on the truncated Normal distribution, which again is not an EDM. Where it is difficult to define the underlying distribution for the performance data being modelled there are numerical methods available that can be used for estimating both the model parameters and their standard errors. These techniques, discussed in Section 3.2.3, employ the bootstrap which involves the creation of empirical distributions that are generated from the sample data itself. Such methods are useful in that accurate estimates for the parameters in any linear model can be obtained.

Simple linear regression models are widely used in the educational context and in many cases researchers fail to assess or report on the underlying assumptions for this model. This project seeks to determine the adequacy of the simple linear regression model if, as is suggested in this chapter, educational achievement data are typically not Normal and contain large quantities of exact zeros. The last section in this chapter detailed techniques for ascertaining this adequacy. In particular, it was hypothesized that educational achievement data in the current context may be generated from a mixture of more definable distributions of data. Any attempts to simulate 'real' educational achievement will therefore rely on such a mixture. Data from each of these defined sub-populations can be generated by one of the probability distributions mentioned in this chapter, and a covariate variable can readily be generated with known linear predictors. The construction and subsequent simulation of data from this model should provide information regarding the adequacy of the simple linear model when it is applied in this particular educational context. The testing and reporting of this simulation will be reported in Chapter 5.

Chapter 4

Results

In the last chapter, the proposed methodology for this study was developed. In particular the primary method of analysis in this study is to be based on the framework encompassed by the theory of generalized linear models. In this chapter these models will be applied, where applicable, to the various forms of performance data generated by students in a tertiary preparatory mathematics class.

The first section of this chapter introduces the data-set used in this study, including the possible explanatory and response variables. The second section reports on the actual modelling of achievement data using the group as a whole, and then using various sub-groups. The last section of this chapter reports on the modelling of progression data.

4.1 Introduction to the data-set

In this study use was made of a data-set that was based on the results of 289 students enrolled in the preparatory mathematics course TPP7181 during 2005. It was obtained from a parallel study aimed to assess the predictive validity of a pre-test that is currently used in the Tertiary Preparatory Program (TPP) at the University of Southern Queensland (USQ). This pre-test,

called the 'M-Test', is composed of four main sections that collectively assess student's knowledge and skills in mathematics, ranging from basic numeracy to calculus. The M-Test has been used by the mathematics team in the TPP for a number of years to assess an appropriate entry level point for commencing students. For a variety of reasons, including the addition of a new mathematics course to the TPP, it was felt that a review of the test's predictive validity should be undertaken. Accordingly all students enrolled in the course TPP7181 in 2005 were required to complete the M-test. Student performance results were then matched with their M-test results.

Of a total enrolment of approximately 700 students, M-test results were available for only 323 students. This was due in part to students failing to complete the test and the TPP administrative team failing to send M-tests to some students. Of the 323 tests available, only 289 students had completed the demographic details that accompany the test, consequently this sub-group together with their final course results formed the basis of the M-test validation data-set. The final set of 289 observations was not randomly obtained, and it is possible that results based on this data-set will be biased. This is a limitation of the current study.

4.1.1 Explanatory variables

Although the discussion in Section 2.2 outlines a number of possible predictors of academic performance, the explanatory variables in this study were limited to those available from the data-set described earlier. These include student results in the M-test and other demographic details, all of which are described in more detail below.

Students are advised at the beginning of the M-test that completion of the first two sections is sufficient for entry into the course TPP7181 (the focus of this study) so while some students attempted later sections of the test, only results in the first two were used. The first section of the test 'Part A', contained 16 items that ranged from basic arithmetic calculations (of the

Highest level	Number
Year 9 or less	20
Year 10	93
Year 11	49
Year 12	113
Other	14

Table 4.1: Highest level of education

type $102 - 36$) to the interpretation of trend graphs, and was scored out of a total of 16. The second section of the test 'Part B', contained 21 items that ranged from concepts of ratio (such as $\frac{3}{4} = \frac{15}{?}$) to drawing a trend graph, and was scored out of a total of 22. Scores in Part A ranged from 4 to 16, while in Part B they ranged from 2 to 22. Total scores ranged from 6 to 38 and the distribution of these totals is shown on the upper graph of Figure 4.1. This distribution is left skewed. Students were also required to provide some demographic data with the M-test and these included: gender; their highest level of schooling; and, the number of years since they last studied mathematics. Of the 289 students in this data-set, 110 were male and 179 female.

Many students had completed year 12 (the senior year in secondary school) although a large number had only completed year 10 (an intermediate exit point in secondary education). A small number of students had studied mathematics at Institutes of Technical and Further Education (TAFE) or even previously in the TPP program. The number of students in each category is shown in Table 4.1.

Most students had studied mathematics formally within the last 5 years, with 37 having studied mathematics within the last 12 months. The distribution of time (in years) since last studied mathematics is shown on the

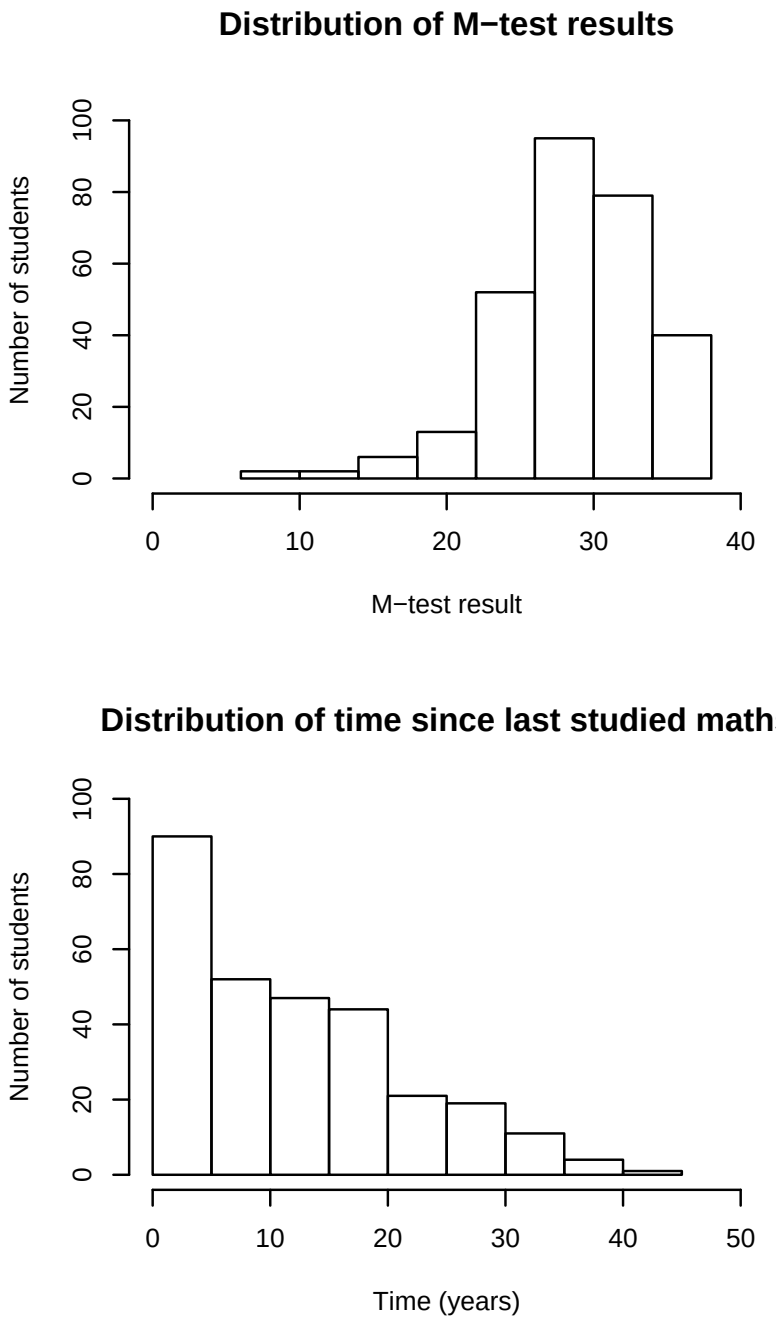

Figure 4.1: Distribution of M-test results (out of 38) shown on top graph and distribution of time since last studied maths on the lower graph

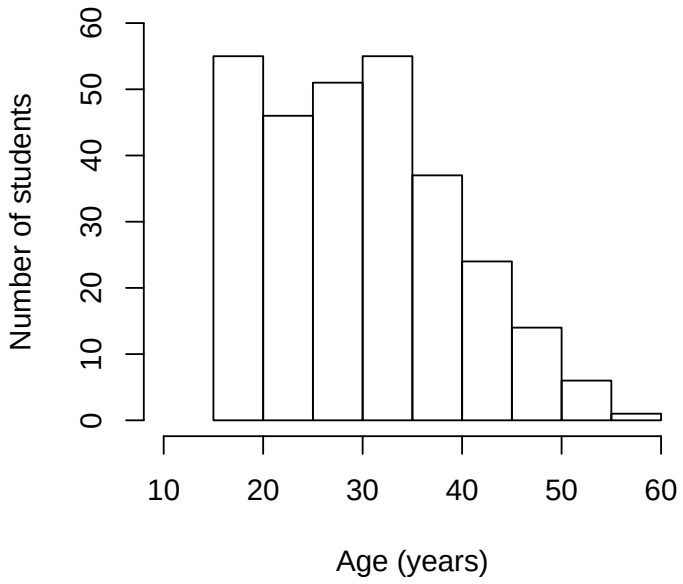

Figure 4.2: Distribution of ages

lower graph of Figure 4.1.

Other demographic data, including the student's age and whether they were currently serving as prisoners were available in the University student records. The average age of students was 30.5 years (the median age was 30 years). Approximately 13% of students were younger than 20 years. The distribution of ages is shown in Figure 4.2. Of the 289 students, 58 were currently prisoners in correctional centres. The vast majority of prisoners (51 of the 58) were males.

Assessment task	Weight (%)
Assignment 1	6
Assignment 2	11
Assignment 3	11
Assignment 4	11
Assignment 5	7
Exam	54
Total score	100

Table 4.2: Assessment tasks and weights for overall course score

4.1.2 Response variables

In Section 2.1 performance data in the educational context were defined to include both measures of student achievement and measures of their progression through the course. For this particular data-set, an aggregated mark (score) for each student served as a measure of both achievement and progression. Students who only completed a few assessment tasks in the course will have a low score, while those who have progressed through the entire course and complete all assessment tasks will generally have a higher score. The number of assessment tasks completed was an alternative measure of progression, as was whether the student actually completed the course or did not complete the course. The score was a weighted aggregate of marks assigned to a number of different assessment tasks. These tasks included five assignments and an examination. The individual weights for these items are shown in Table 4.2. The distribution of student scores (as a percentage) is shown in the bottom right hand plot of Figure 4.3 which also shows the distribution for each of the student sub-groups defined in Section 3.4. The distribution for the entire group is obviously non-symmetric, in fact bi-modal, but as discussed in Section 3.4 it can be thought of as being a mix-

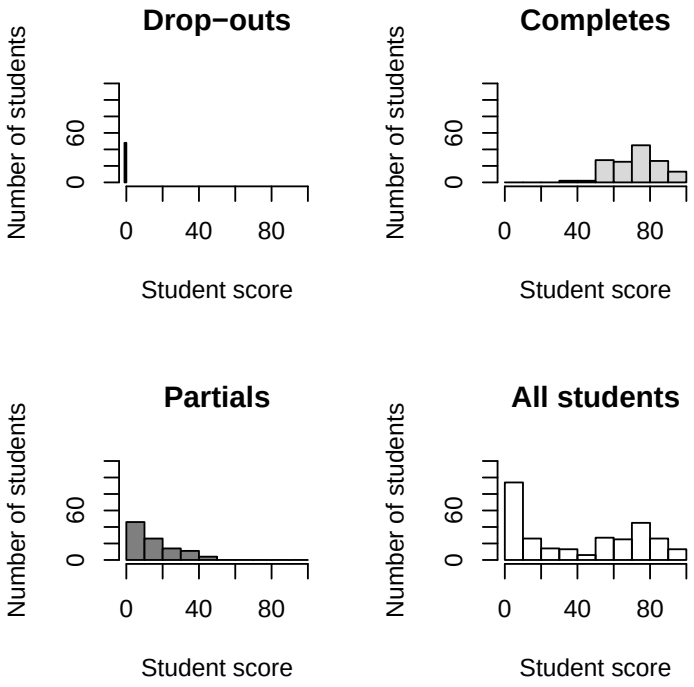

Figure 4.3: Distribution of student score by population sub-group

ture of three distinct distributions (shown in the remaining plots of Figure 4.3). In this particular data-set, 48 students (17%) withdrew from the course without submitting any piece of assessment and were subsequently given an exact zero (classified as 'drop-outs'). Of the remaining students, 101 (35%) failed to complete the course and in particular did not sit for the final course examination which contributed 54% towards the total (these students were classified as 'partials').

The number of assessment tasks completed by the students ranged from 0 to 6 (see Table 4.3). There was a total of 136 students who completed all requirements of the course. An additional four students completed the examination but did not complete one assignment. These students can be

Total number of items	Number of students
0	48
1	43
2	25
3	13
4	7
5	13
6	136

Table 4.3: Total number of assessment items completed

regarded as having completed the course. Consequently, for modelling purposes, there were 140 complete students (48%) and 149 incomplete students (52%) for this course.

4.2 Modelling achievement results

In the last section the actual data-set used in this analysis was introduced. It was noted that the student score could be regarded as being both a measure of student achievement of the course objectives and a measure of their progression through the course. In this section the variable score is used to model student achievement. It should be noted, however, that this variable is very much dependent on how long the student remains in the course. Incomplete students will not have been exposed to material that addresses some or all of the course learning objectives. For this reason it may be more appropriate to model score for the two major sub-groups of students, namely those who complete the course and those who do not.

4.2.1 Modelling achievement of complete students

As discussed in Chapter 2, most studies dealing with modelling student academic achievement examine only students who complete the course; after all, it is assumed that the remainder, for varying reasons, were unable to complete the course and fully demonstrate their achievement. Achievement data for students who completed the course TPP7181 were modelled using the set of explanatory variables introduced in the last section. The results of this modelling are discussed in this section.

There were 140 students in the data-set that fully completed the course. The distribution of their total mark is reasonably symmetric and is shown in the top right plot of Figure 4.3. This would suggest that standard linear regression might be an appropriate method for modelling student achievement. Alternatively it might be possible to model student achievement using a linear model that is based upon the beta distribution. As discussed in Section 3.2.3, beta regression is useful for modelling proportions (in this case the total expressed as a proportion). Moreover the beta distribution can accommodate a range of non-Normal distributions. In this section both models were applied to these data and the results are discussed below.

Linear model

The strength of the linear association (as measured by the Pearson correlation coefficient) was calculated for each pair of continuous variables in the data-set. The strongest correlation of 0.49 existed between the student score (Score) and the M-test result (Mtest) and indicates that this explanatory variable should be included in any linear model. A scatterplot of these two variables is shown in Figure 4.4, which includes a local fitted polynomial regression curve. This curve is obtained through a process developed by Cleveland & Devlin [25] that fits a polynomial to a small sample of points near a given value of the explanatory variable, say $x = x_i$. This fitted polynomial, in turn, is used to predict the value of the response for $x = x_i$, and the

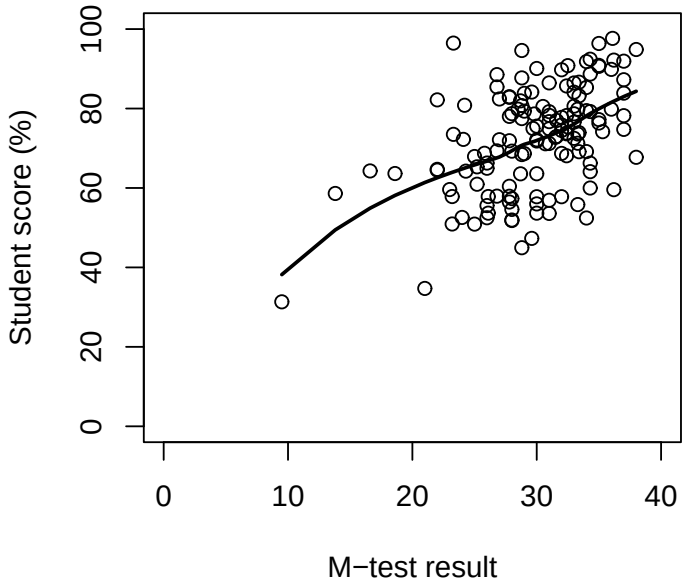

Figure 4.4: Student score against M-test result for complete students, with local fitted polynomial regression curve.

process is repeated for all values of x. From this plot it can be seen that the polynomial is approximately linear, which supports the appropriateness of a linear model in this situation.

An initial linear model was fitted to the data-set and was able to explain 33.6% of the variation in performance scores. Statistically significant explanatory variables (at the 5% level) in the data-set were: M-test result ($p \approx 0$), the gender of the student ($p = 0.01$) and the number of years since the student last studied ($p \approx 0$). This model is shown in Equation (4.1) where the variable M-test result is denoted 'Mtest' and the variable years since last studied 'Time'. The equation also includes standard errors

of estimates, which are placed in brackets underneath each estimate.

$$\underset{}{\text{Score} = } \underset{(6.2)}{21} + \underset{(0.20)}{1.5}\ \text{Mtest} + \underset{(1.9)}{4.8}\ \text{Gender} + \underset{(0.10)}{0.38}\ \text{Time} \tag{4.1}$$

In Equation(4.1) the variable Gender is defined:

$$\text{Gender} = \begin{cases} 0 & \text{if Male} \\ 1 & \text{if Female.} \end{cases}$$

In this model an increase in the M-test of 1 mark, with all other factors remaining constant, will lead to an average increase in student score of 1.5 percentage points. Similarly, with all other factors remaining constant, female students will on average have a score 4.8 percentage points higher than males. Every additional year since last studied mathematics will, on average and with all other factors remaining constant, lead to an increase in student score of 0.38 percentage points. Such conclusions, however, need to be considered in light of the particular context, for as MacGillivray & Turner [60, p. 78] point out 'analyzing student results requires considerable care because of the many possible interdependencies and confounding or hidden variables ...'.

A statistically significant interaction was found to exist between scores in the M-Test and the semester in which the student studied. A further linear model (see Equation (4.2)) was created that was able to explain 35% of the variance in student scores. The variable semester ($p = 0.04$) and the interaction term ($p = 0.05$) were both statistically significant at the 5% level (the constant in this model was not statistically significant, $p = 0.7$). Polynomial functions of each of the continuous explanatory variables were also tested for statistical significance, however in each case only the linear

term was significant.

$$\begin{aligned} \text{Score} \;=\; & \underset{(10.1)}{3.2} \;+\; \underset{(0.3)}{2.0}\,\text{Mtest} + \underset{(1.9)}{4.7}\,\text{Gender} \\ & + \underset{(0.1)}{0.40}\,\text{Time} + \underset{(12.2)}{25.9}\,\text{Semester} - \underset{(0.4)}{0.80}\,\text{Mtest}\times\,\text{Semester} \end{aligned} \tag{4.2}$$

In this case the variable Gender is defined as above and the variable Semester is defined:

$$\text{Semester} = \begin{cases} 0 & \text{if enrolled in semester 1} \\ 1 & \text{if enrolled in semester 2.} \end{cases}$$

The large effect estimate for Semester ($\beta = 25.9$) in this model, may be explained by the existence of a few influential observations (see Figure 4.5). In this figure influential points (numbered) are identified through large values of the corresponding hat matrix. Point 1 from semester 1 and points 2 and 5 from semester 2 in a sense pull the lines apart, creating the large effect difference. The interaction term between M-test result and semester may simply be a result of two different markers used in the two semesters. A comparison of the distribution of M-test results for both semesters is shown in Figure 4.6 and indicates that in general, semester 2 M-test results were lower than those obtained in semester 1. There were slightly more females in the semester 2 cohort, so this may also have contributed to the interaction term, as generally females appear to perform better than males. The mean time since last studied and student score were similar for both semester groups. A plot of the residuals against the major explanatory variable (M-test results) does not indicate any violation in the assumption that the variance is constant (see the top graph in Figure 4.7). Similarly a plot of residual quantiles against theoretical quantiles does not indicate a violation in the assumption that the residuals are distributed Normally (see the lower graph in Figure 4.7), although there is a slight departure for larger values. The five points that were identified as being influential are also shown on the top

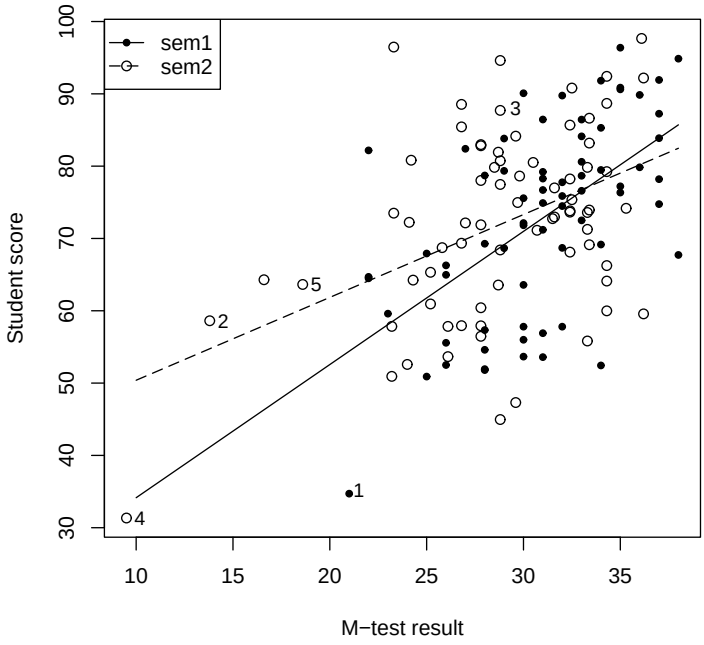

Figure 4.5: Student score against M-test result by semester, with simple linear models of score against M-test result shown for each semester.

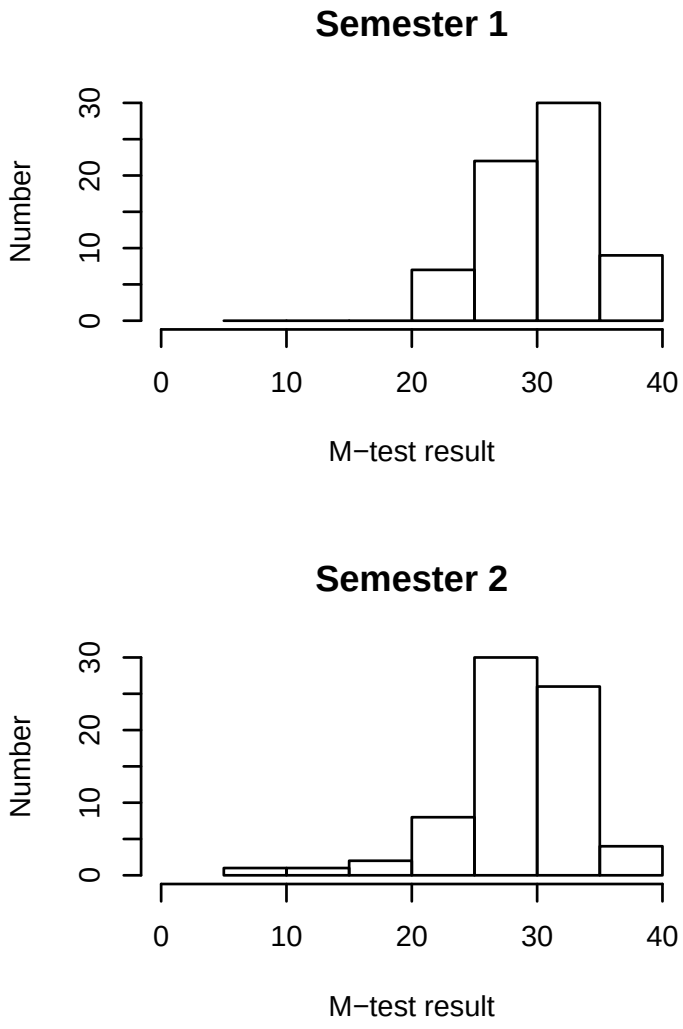

Figure 4.6: Distribution of M-test results by semester

graph of Figure 4.7. Of the two most influential points, one (labelled 2 on the graph) represents a student who scored a very low pre-test result yet passed the course. The other (labelled 4 on the graph) belongs to the only student who completed all assessment items and failed the course.

The linear model, described in Equation(4.2) was able to explain 35% of the variation, which is comparable with models reported in the literature (see Robbins et al. [84]). The model, however, may be inappropriate for these data. The spread of residuals shown in the top graph of Figure 4.7, and the slight departure from linearity, of the quantile–quantile plot (shown in the lower graph of Figure 4.7), point to a skewness in the data. Moreover the existence of outliers may indicate that a non-linear model is more appropriate. For these reasons a model based upon the beta distribution will be considered for these data.

Beta model

In Section 3.2.3 the beta regression was introduced as a possible method for dealing with non-Normal data. It was noted that a linear model based upon the beta distribution can better represent the data that in this instance have fixed upper (100%) and lower (0%) limits. For this reason and those mentioned earlier, a model based upon the beta distribution (see Equation (4.3)) was also applied to this data-set and was able to explain 36.1% of the variation in performance (this figure is based on a pseudo-R^2 measure defined by Ferrari & Cribari-Neto [38]). In this model the statistically significant explanatory variables at the 5% level were: the M-test result ($p \approx 0$), the number of years since the student last studied ($p \approx 0$), the gender of the student ($p = 0.02$), the semester in which the course was studied ($p = 0.02$)

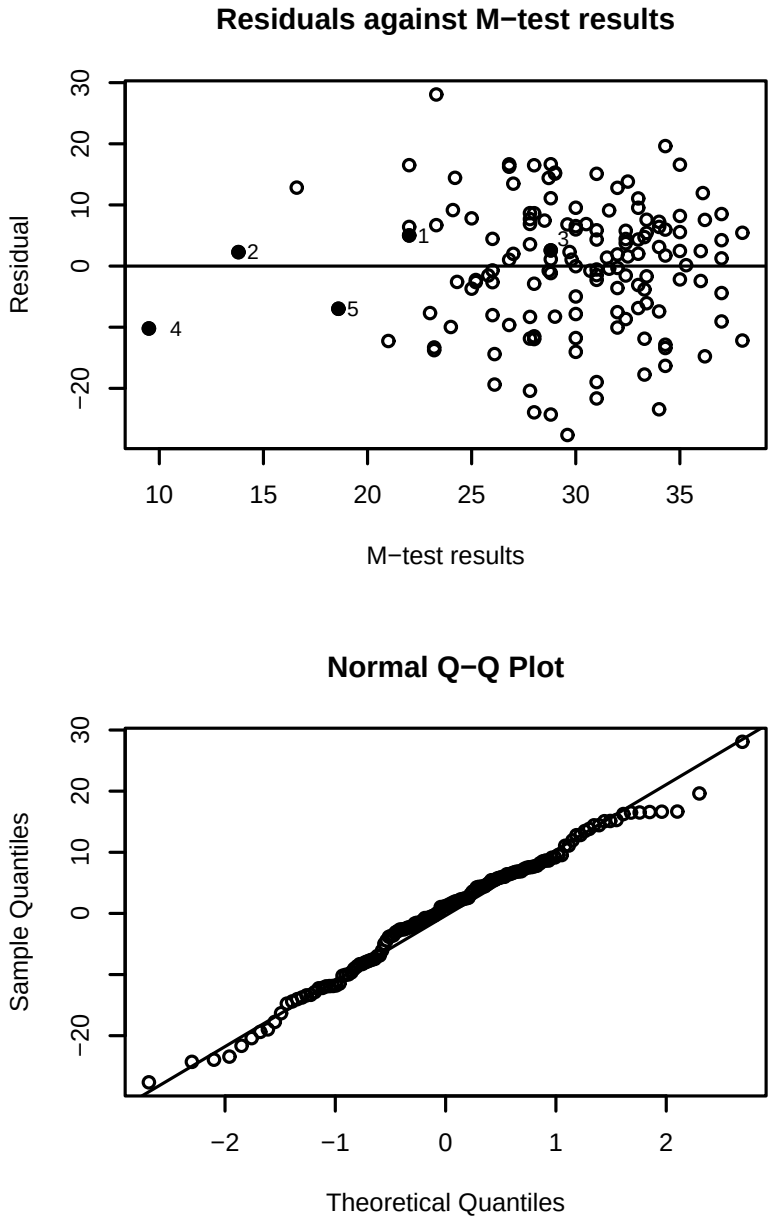

Figure 4.7: Plot of residuals against M-test results for the standard linear regression model (influential points numbered 1 to 5) shown on the top graph. Plot of empirical quantiles against theoretical quantiles shown on the lower graph.

and the interaction between semester and the M-test result ($p = 0.03$).

$$\begin{aligned}\text{logit(Score)} &= \underset{(0.5)}{-2.5} + \underset{(0.01)}{0.1}\ \text{Mtest} + \underset{(0.1)}{0.2}\ \text{Gender} + \underset{(0.01)}{0.02}\ \text{Time} \\ &\quad + \underset{(0.6)}{1.4}\ \text{Semester} - \underset{(0.02)}{0.04}\ \text{Mtest}\times\text{Semester} \end{aligned} \tag{4.3}$$

If the linear predictor of the above model (that is the right hand side of equation 4.3) is denoted as $\boldsymbol{x}_i^T\boldsymbol{\beta}$, then the model can be written:

$$\text{Score} = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} \tag{4.4}$$

The linear model (Equation (4.2)) and the beta model (Equation (4.3)) were similar in that they identified the same explanatory variables, and were able to account for approximately the same amount of variation in student scores. The similarity between the residual plot for this model (shown in Figure 4.8) and that for the linear model (see Figure 4.7) also indicates that the two models were comparable. Whilst models that are based upon the beta distribution are difficult to interpret, the logit transformation of the response allows for a non-constant change in the effects ($\boldsymbol{\beta}$) over the domain of the explanatory variables ($\boldsymbol{x}_i$). As is reported in Chapter 5, the assumption of a constant effect (as is the case for the standard linear regression model) may produce inaccurate predictions at either end of the explanatory variable domain.

4.2.2 Modelling achievement of incomplete students

Of interest in this study is whether in fact the 149 observations belonging to those students who did not complete the course can be used. One of the major problems with this subset of the data is the inclusion of 48 zeros for students who withdrew from the course without having submitted any assessment task. In Section 3.3 it was noted that both the Tweedie model and the Tobit

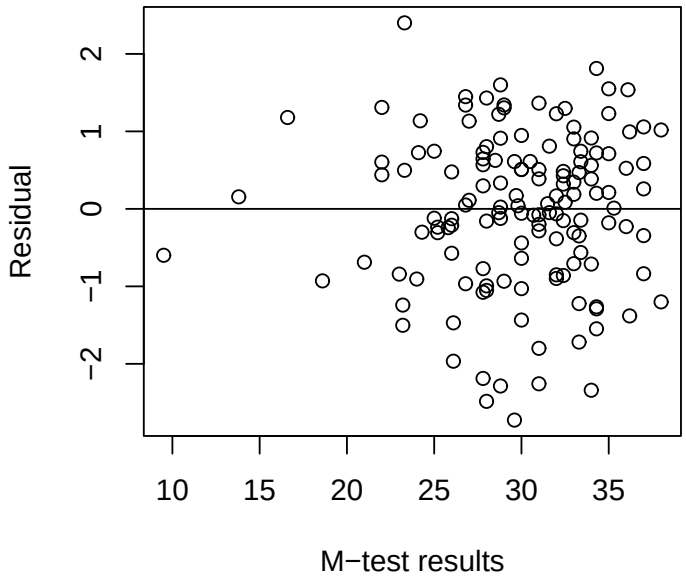

Figure 4.8: Plot of residuals against M-test results for the beta regression model.

model may be appropriate for modelling educational performance data. Both of these models were applied to the achievement data of those students who did not complete the course TPP7181 and the results are discussed below.

The Tweedie model

As detailed in Section 3.3.3, the Tweedie model is a generalized linear model whose response variable is distributed according to a Tweedie exponential dispersion model. This section reports the results of fitting such a model to the performance data of students who did not complete the course.

The distribution of scores for the 149 incomplete students is shown in Figure 4.9, which includes the 48 students who dropped from the course without ever providing any performance measure. Maximum likelihood estimation of the Tweedie p-parameter was undertaken using the R-package `tweedie` [32]. A Tweedie distribution with parameter $p = 1.087$ is also shown on Figure 4.9.

A generalized linear model, based upon this distribution and with a log link was then fitted to the scores for these students. Of the possible explanatory variables available in this data-set, only students results in the M-test ($p \approx 0$) and whether they had completed school at a higher level than Year 10 or not ($p \approx 0$), which will be termed 'Junior', were statistically significant. The constant term, however, was not statistically significant ($p = 0.26$). A significant interaction was also noted between the variables Junior and Mtest ($p \approx 0$). The final model, shown in Equation (4.5), had a pseudo-R^2 value (as calculated using Equation (3.11)) of 8.8%. Such a low value of R^2 indicates that the model is quite inadequate in explaining the considerable variation in scores for these students.

$$\begin{aligned} \log(\text{Score}) &= \underset{(0.96)}{-1.08} + \underset{(0.03)}{0.12}\ \text{Mtest} + \underset{(1.16)}{3.27}\ \text{Junior} \\ &\quad - \underset{(0.04)}{0.11}\ \text{Mtest}\times\text{Junior} \end{aligned} \tag{4.5}$$

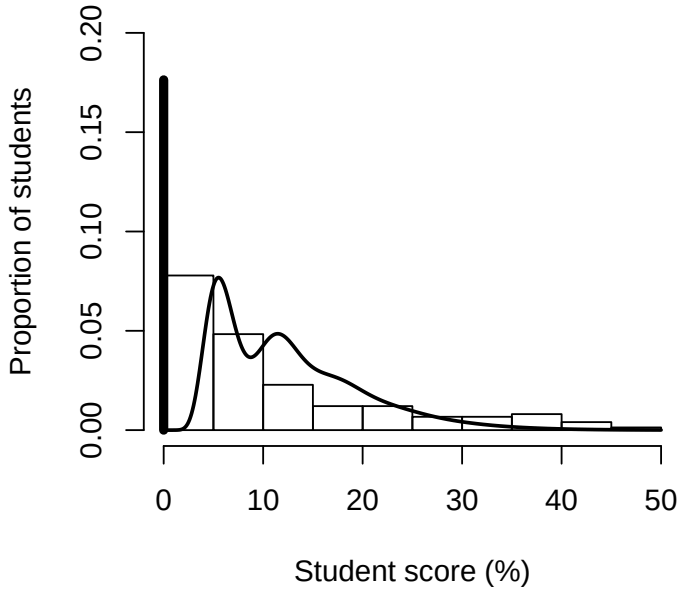

Figure 4.9: Distribution of student scores for incomplete students with Tweedie ($p = 1.07$) distribution also shown

The variable Junior in this equation is defined:

$$\text{Junior} = \begin{cases} 0 & \text{if student has not completed school to a Year 10 level} \\ 1 & \text{if student has completed school to a Year 10 or higher level.} \end{cases}$$

The model shows that with all other factors constant, each extra mark in the M-test will lead to an increase in the logarithm of the student score of 0.12, for students who have not completed junior. For students who have completed junior, however, the effect of the M-test score on final performance is reduced considerably. This result is not surprising, as the material assessed in the M-test is normally covered in the early secondary school years. Students who have remained in school beyond a year 10 level should have been repeatedly exposed to this material. Consequently the M-test results appear to be a reasonable measure of prior-knowledge for students who have not completed junior but inadequate for those who have. What is surprising is that this interaction between junior and M-test is not evident for students who have completed the course.

Five influential points (shown on the top graph of Figure 4.10) were identified. Omitting these points from the model made little difference to model estimates but did achieve a slight improvement in the pseudo-R^2 value. Of the five influential points shown, the most influential (numbered 3) represents a student who scored the lowest in the M-test, yet had completed Year 10 and was able to achieve a total score of 21.5%.

A plot of quantile residuals for this model against M-test results is also shown on the top graph of Figure 4.10. This plot does not provide any evidence to suggest that the linear model in this instance is inappropriate. A plot of quantile residuals against theoretical quantiles is shown on the lower graph of Figure 4.10. This plot does not provide any evidence suggesting an inappropriate choice of distribution and link function.

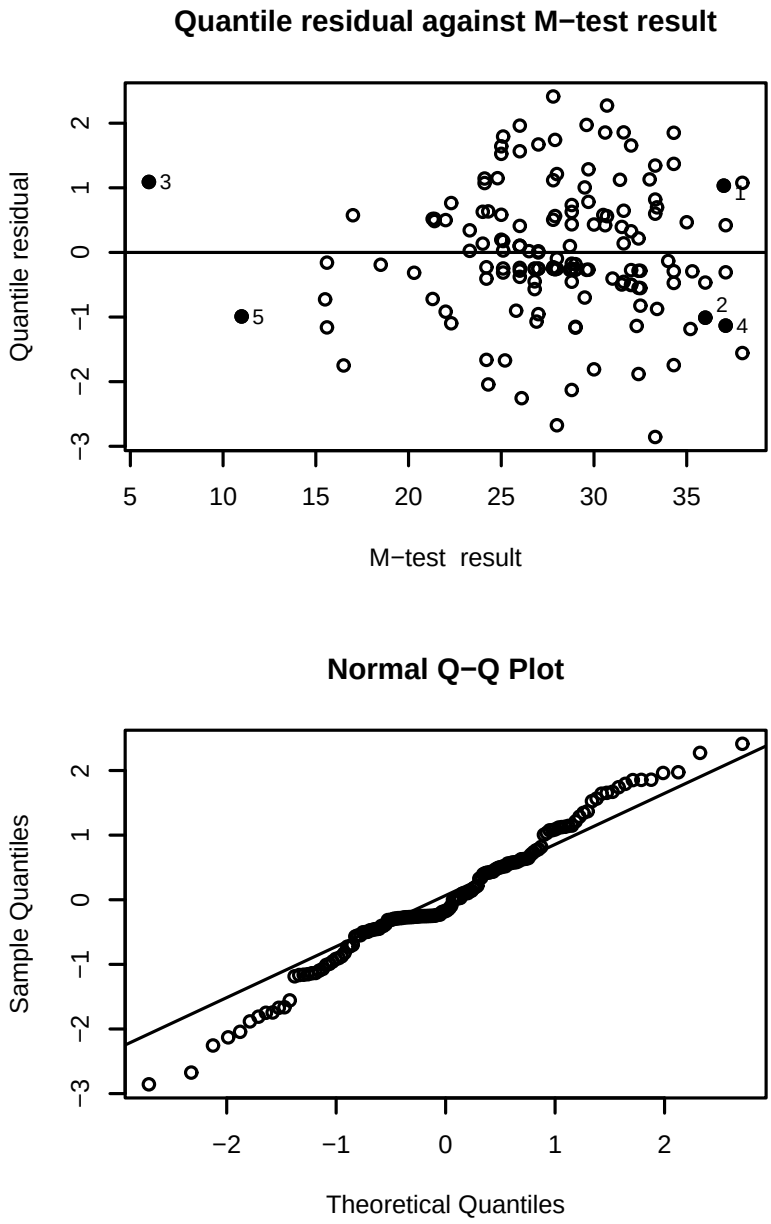

Figure 4.10: Plot of quantile residuals against M-test results for the Tweedie model (influential points are labelled 1 to 5) shown on the top graph. Plot of residual quantiles against theoretical quantiles shown on the lower graph.

Tobit model

As detailed in Section 3.3.2, the Tobit model assumes the existence of an underlying latent performance variable. Further, it is assumed that this performance variable can only be measured when the amount that students possess exceeds some estimated threshold. The model is then fitted to the existing achievement data on the basis that the exact zeros represent performance results for students who have not exceeded the threshold.

The Tobit model was fitted to this data-set using the R-package `survival` [75]. At the 5% level of significance, only results in the M-test explained the variable student score ($p = 0.01$). The model in this instance is:

$$\text{Score} = \begin{cases} 0 & \text{if Mtest} \leq 18.8 \\ 0.73 \times \text{Mtest} & \text{if Mtest} > 18.8 \end{cases} \tag{4.6}$$

The Tobit model, in this instance, predicts a zero performance when student M-test results fall below 18.8 marks. Of the 8 students in this category, 6 obtained a zero performance score. There were, however, 40 students with M-test results greater than 18.8 who obtained a zero performance score. Consequently the predictive ability of this model appears to be limited. Further it is probably inappropriate to calculate a pseudo-R^2 value for this model, as the model assumes the existence of an underlying latent explanatory variable.

A graph of the above model is shown in Figure 4.11, which also shows the Tweedie model for the same data. The later was obtained by fitting a Tweedie linear model with just M-test results as the explanatory variable (pseudo-R^2 value of 4.2%). The exponential curve produced by the Tweedie model appears to better summarize variation in the data than the two step linear function of the Tobit model, although the later is of course far more accurate for M-test values less than 18.8. Due to the obvious lack of suitable explanatory variables, both models do not explain very much of the considerable variation in student achievement. With more research in this particular

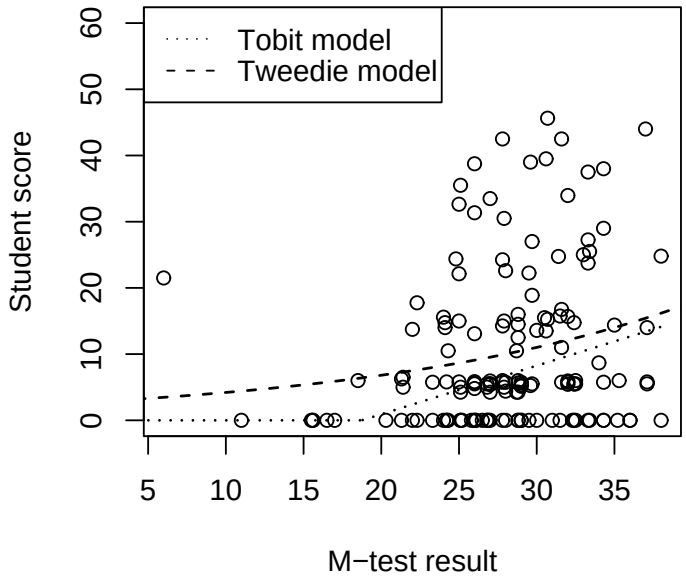

Figure 4.11: Comparison of Tweedie and Tobit models

area aimed to identify such explanatory variables, it is conceivable that the two models could be more effective.

4.2.3 Modelling achievement of all students

In this section, the achievement variable 'score' has been modelled for the two major sub-groups of students, namely those who complete the course and those who do not. An inspection of the distribution of scores for the entire group (see Figure 4.4) provides support for the view that the total distribution is in fact a mixture of results from these two sub-groups. Nevertheless, and despite the obvious lack of Normality, a standard linear regression model was fitted to this variable for the entire group of students. The reason for

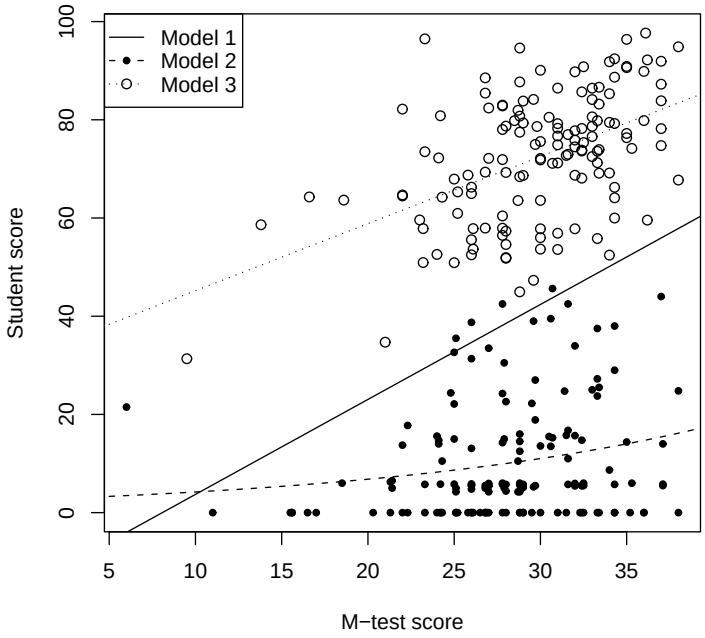

Figure 4.12: Models of achievement for all students and sub-groups, with M-test results used as the explanatory variable. Model 1 is a linear model applied to all students; Model 2, a Tweedie model applied to incomplete students; and, Model 3 a linear model applied to complete students.

this was to use this model as a comparison to the other models that have been discussed in this section. Many studies that fit models to educational performance data fail to report on violations of Normality, such as in this instance. It is informative to see how the results might differ if in fact the obvious non-Normality in the data was ignored.

The linear model in question was fitted to the score for the entire cohort of students using the dominant explanatory variable, that is the M-test result. The model (see Equation (4.7)) not surprisingly only explained 8.4% of the variation and is shown as Model 1 on Figure 4.12.

$$\begin{array}{rcrr} \text{Score} & = & -15.6 + & 1.9\ \text{Mtest.} \\ & & (11.0) & (0.4) \end{array} \tag{4.7}$$

In line with the previous treatment of performance data for incomplete students, a Tweedie model was fitted to the scores for these students (shown on the graph as Model 2). This model, see Equation (4.8) had a pseudo-R^2 value of 4%.

$$\begin{array}{rcrr} \log(\text{Score}) & = & 0.95 + & 0.05\ \text{Mtest.} \\ & & (0.6) & (0.02) \end{array} \tag{4.8}$$

Figure 4.12 also shows a linear model fitted to the score of only students who completed the course (shown as Model 3). This model (see Equation (4.9)) is able to explain 24% of the variation in student scores for the sub-group in question.

$$\begin{array}{rcrr} \text{Score} & = & 31.5 + & 1.3\ \text{Mtest.} \\ & & (6.3) & (0.2) \end{array} \tag{4.9}$$

Figure 4.12 provides a visual confirmation that if it is the intention to use regression models to predict student performance, then the most appropriate way of dealing with these data would be to treat them as two sub-groups, that is those students who complete the course and those that do not. It should be noted, though, that the effect size for Model 1 ($\beta_1 = 1.9, \sigma_\beta = 0.4$) was not significantly different from the effect size given in Model 3 ($\beta_1 = 1.3, \sigma_\beta = 0.2$). In other words, if it is the intention to estimate the effect of prior knowledge on student performance there is little to be gained, in this instance, from taking care with model assumptions (in this instance Normality). Whether such a conclusion can be generalized to all educational performance data will be explored further in the following chapter.

4.2.4 Summary

In this section achievement data for students were modelled using a number of approaches. In the first instance, data for those students who completed the course were modelled using a standard linear regression model and a model based upon the beta distribution. Residual plots indicated that both of these models were reasonably appropriate in this instance. Data for those students who did not complete the course were then modelled through the application of both Tweedie and Tobit models. Both of these models have a certain intrinsic appeal in that they attempt to explain the considerable number of students who do not complete the course and were assigned zero. In both models, there was insufficient information available to adequately explain the variation in performance for students who do not complete the course. In the last part of this section a linear model was applied to the achievement data for the entire group of students. It was found that if the intention of the model was to predict student achievement, then such a treatment would not be desirable. If the intention for the model was to ascertain the effect size, then ignoring obvious violations in Normality and applying a linear model to the entire group is in reality no worse a method than applying a linear model to the data for students who complete the course.

4.3 Modelling progression

In the last section the modelling of achievement data for students in the course TPP7181 was discussed. In particular it was noted that the modelling of such data should be based on sub-groups of the data-set. In this section the modelling of alternative measures of progression is undertaken.

4.3.1 Modelling completeness

The most commonly used measure of progression is whether a student actually completes the course or not. In this study a dichotomous variable was created in order to model a student's ability to complete the course. Of the 289 students in the data-set, 140 completed the course and 149 did not. A generalized linear model was fitted to this dichotomous variable using the binomial distribution as the underlying distribution and the logit link. As mentioned earlier, the logit (natural logarithm of the odds ratio) transformation of the response variable, which is defined on the interval $[0, 1]$, ensures that the transformed data are defined for all real values. The details of this model are reported in this section.

Each of the explanatory variables described in Section 4.1.1 from the data-set were tested for significance. The variables that were found to be statistically significant (at the 5% level) were: the result in Part B of the M-test ($p \approx 0$), termed 'Mtest(B)', and whether the student had completed Year 12 prior to enrolling in the course ($p = 0.05$). The later was a dichotomous variable created from information given by students regarding their highest year level studied and called 'Senior'. A significant interaction term (at the 5% level) was also observed between Mtest(B) and Time ($p = 0.02$). The pseudo-R^2 value for the model (calculated according to Equation 3.11) was only 6.3%, in other words the model was not able to explain very much of the variation in performance for this group of students. The model in this instance is:

$$\begin{aligned} \text{logit(Complete)} &= \underset{(0.8)}{-2.5} + \underset{(0.05)}{0.2}\ \text{Mtest(B)} + \underset{(0.05)}{0.08}\ \text{Time} \\ &\quad + \underset{(0.3)}{0.5}\ \text{Senior} - \underset{(0.02)}{0.01}\ \text{Mtest(B)}\times\text{Time}, \end{aligned} \tag{4.10}$$

where Complete is defined:

$$\text{Complete} = \begin{cases} 0 & \text{if student has not completed the course} \\ 1 & \text{if student has completed the course,} \end{cases}$$

and Senior is defined:

$$\text{Senior} = \begin{cases} 0 & \text{if student has not completed Year 12 at school} \\ 1 & \text{if student has completed Year 12.} \end{cases}$$

In this model a student's educational background, in this case whether they have completed senior or not, will contribute the most to the probability of their completing the course. The statistically significant interaction term in this model indicates that the effect of prior knowledge (as measured in Part (B) of the M-test) is diminished by the length of time since the student last studied. Arguably other factors such as family commitments may play a more major role in determining whether students complete the course for more mature students. With such a large amount of unexplained variance this interpretation is merely speculative, however it does support similar findings reported by Petrides et al. [79].

4.3.2 Modelling type

An alternative to dichotomizing the complete status of a student is to create an ordinal variable (Type) with three categories, namely those students who drop-out of the course without providing any performance data (Type = 1), those who continue with the course but do not complete the examination (Type = 2) and those who complete all components of the course (Type = 3). As discussed in Section 3.2.2 an ordinal regression model can be fitted to this variable. In this section the results of such a process are reported.

An ordinal model was applied to the data-set using the R-package `MASS`

[97]. Of the suitable explanatory variables in the data-set, the following were found to predict the type of student: the student's result in the M-test ($p \approx 0$), their age in years ($p \approx 0$), which semester they studied the course ($p \approx 0$) and whether they completed Year 12 at school ($p = 0.02$). These variables were all statistically significant at the 5% level and their was no significant interaction term. The model is shown in Equations (4.11) and (4.12) and has a pseudo-R^2 of 6.9%.

$$\begin{aligned} \text{logit}(P(\text{Type} \leq 1)) &= \underset{(0.8)}{-1.1} \underset{(0.02)}{-\ 0.08}\ \text{Mtest} \underset{(0.01)}{+\ 0.04}\ \text{Age} \\ &\quad \underset{(0.3)}{+\ 0.8}\ \text{Sem2} \underset{(0.2)}{-\ 0.5}\ \text{Senior} \end{aligned} \tag{4.11}$$

$$\begin{aligned} \text{logit}(P(\text{Type} \leq 2)) &= \underset{(0.9)}{0.8} \underset{(0.02)}{-\ 0.08}\ \text{Mtest} \underset{(0.01)}{+\ 0.04}\ \text{Age} \\ &\quad \underset{(0.3)}{+\ 0.8}\ \text{Sem2} \underset{(0.2)}{-\ 0.5}\ \text{Senior} \end{aligned} \tag{4.12}$$

From this model it can be seen that similar to the logistic model described in the last section, a student's highest educational achievement contributes the greatest to the probability that they will either drop from the course or fail to complete the course. Their result in the M-test (rather than in the second part of the M-test) will also explain this. The higher constant term in Equation (4.12) merely indicates that the probability that students will be of Type = 1 (drop-out) or Type = 2 (partial), is greater than the probability that they are only of Type = 1.

As discussed in Section 4.1.2, an alternative variable for modelling progression is the number of assessment tasks that the student completes. An ordinal model was applied to the data-set with the number of assessment tasks as the response. This model was only able to explain 3.2% of the variation and consequently will not be reported further.

4.3.3 Summary

In this section various measures of student performance that related to student progression were modelled. In all models, the set of explanatory variables available, were not sufficiently able to explain the variation in progression measures. In other words, those variables that explained achievement were not as effective in explaining progression. This result is not unusual, with De Berard et al. [28] able to produce a model of student achievement but unable to produce a model of student retention using the same set of predictor variables. Arguably, it is not a worthwhile exercise to evaluate the merits or otherwise of the models outlined in this section. More research needs to be undertaken in order to identify and measure suitable predictors of student progression through the course, before such an evaluation occurs.

4.4 Model Validation

In this chapter many types of models have been applied to educational performance data. The utility of such models has often been evaluated according to a measure of how much variation in performance the respective model can explain. An alternative approach to use, especially when the model is to be used for predictive purposes, is to validate the model using an independent data-set. In this section, some of the models developed earlier in this chapter are applied to an alternative, but similar data-set.

It is reasonable to assume that student characteristics remain reasonably constant from one semester offering of an educational course to the next, so that performance data from a subsequent semester of the course TPP7181 should serve as suitable validation data. The statistically significant influence of the variable semester on performance, as reported in Equations (4.2) and (4.3) was attributable to one or two influential points, which may be unlikely to occur in subsequent semesters. The data used in this section, and subsequently referred to as the validation data-set, were obtained from

student results in the course TPP7181 in semester 1 of 2006. The validation data-set contained 179 observations and all of the explanatory variables that were used in the original data-set except for semester (the data in this instance was obtained from only one semester). Consequently those models that did not include the semester of study as an explanatory variable were fitted to the data-set, and the results are detailed below.

4.4.1 Validation of models of achievement

In Section 4.2 a number of models were applied to the achievement of different sub-groups of students. Osborne [76] recommends that the cross-validation-coefficient (CVC) is a suitable measure for evaluating models of continuous data. In order to evaluate this coefficient for a model, it is applied to an independent sample of data and predicted values are calculated. The CVC is then the degree of correlation (Pearson's correlation coefficient) between these predicted values and the observed values for the independent sample. The difference between the square of the CVC and the R-square of the original model (called the shrinkage) should be low, although there are no guidelines as to how low.

The standard linear regression model shown in Equation (4.1) was applied to the performance data of students in the test data-set who completed the course ($n = 97$). The CVC in this instance was calculated to be 0.43, and the shrinkage, a 15 percentage point difference. This is quite a large reduction in explained variation and suggests that the model in question is poor in its ability to generalize beyond the particular data from which it was created.

In order to validate the beta model shown in Equation (4.3), an alternative model that did not include the variable semester was used. Each of the variables in the model (shown in Equation (4.13)) were statistically

significant (at the 5% level) and the model had a pseudo-R^2 value of 33.2%.

$$\begin{aligned} \text{logit(Score)} &= \underset{(0.31)}{-1.5} + \underset{(0.010)}{0.070}\ \text{Mtest} + \underset{(0.10)}{0.23}\ \text{Gender} \\ &\quad + \underset{(0.0052)}{0.022}\ \text{Time} \end{aligned} \tag{4.13}$$

This model was applied to the performance results of complete students in the test data-set. The transformation shown as Equation (4.4) was then applied to these predicted values. Although the CVC is a measure usually applied to standard linear regression models, it was applied to this model also, in order to provide a means of comparing the linear and beta models. The CVC in this instance was calculated to be 0.44, which suggests that the beta and linear models are equally effective. An estimate of the shrinkage based on the pseudo-R^2 value of the beta model shown in Equation (4.13) was a 14 percentage point difference, which again shows that the linear and beta models are comparable.

The Tweedie regression model shown in Equation (4.5) was then applied to the performance data of students who did not complete the course ($n = 82$). Predicted performance scores failed to correlate with observed performance scores, with a CVC of 0.14. One of the problems in this instance may have been the 30% reduction in the proportion of exact zeros, 17% in the original data-set to 12% in the test data-set. In any case, this result points to a major limitation of this particular Tweedie model in explaining the performance of students. The estimated shrinkage of the Tweedie model, based on the reported pseudo-R^2 of 8.8%, was approximately 6.8% in real terms (77% as a percentage of the original pseudo-R^2 value). The Tobit model shown in Equation (4.6) was also applied to the performance data of incomplete students in the test data-set. It produced a CVC of 0.08, an estimate of the shrinkage was not obtained as a pseudo-R^2 measure was not made for the Tobit model. This low CVC value suggests that the Tobit model is less useful than the Tweedie model when used for predictive purposes.

4.4.2 Validation of models of progression

The two models discussed in Section 4.3 both predicted probabilities of students obtaining given outcomes. Kotsiantis et al. [53] define two measures for evaluating such predictive models:

1. The sensitivity of a model is how effective the model is in correctly classifying positive outcomes, so in the educational context a positive outcome might be that the student obtains a given grade. The sensitivity of a model is defined as:

$$\text{sensitivity} \quad = \quad \frac{a}{a+b}, \tag{4.14}$$

   where a is the number of positive outcomes correctly predicted and b the number of positive outcomes incorrectly predicted as being negative.

2. The specificity of the model is how effective it is in correctly predicting negative outcomes. Again, in the educational context, a negative outcome might be that the student does not obtain a given grade. The specificity is defined as:

$$\text{specificity} \quad = \quad \frac{d}{c+d}, \tag{4.15}$$

   where d is the number of negative outcomes correctly predicted and c the number of negative outcomes incorrectly predicted as being positive.

Arguably, an effective model should have measures of sensitivity and specificity close to 1. Both of these measures were used to evaluate the models of progression in this chapter.

The model of completeness, shown as Equation (4.10), was applied to the test data-set and predicted probabilities of completion were calculated. If the probability was bench-marked, so that a probability of completion in excess

	Predicted drops	Predicted partials	Predicted completes	Total
Observed drops	0	6	4	10
Observed partials	1	19	52	72
Observed completes	2	12	83	97
Total	3	37	139	179

Table 4.4: Cross tabulation of predicted counts against observed counts for the ordinal model of progression

of 0.5 will predict completion, then the two measures described above can be calculated. In this instance the sensitivity was 47% and the specificity 63%. In other words, the model seems to be able to predict negative outcomes more frequently. It should be noted that the sensitivity of the model can be improved if the arbitrary benchmark chosen above is decreased.

The model of student type, shown in Equations (4.11) and (4.12), was also applied to the test data-set. This model is able to calculate the probability that a given student is in any of three category types (drop, partial or complete). It can then be assumed that a student's predicted type will be that category which has the highest probability. Table 4.4 shows a cross-tabulation of the observed and predicted counts for each category when the model was applied to the test data-set. Applying the measures used earlier to this, the sensitivity of the model in predicting that a student completes the course is 86%: it correctly identifies 86% of all students who complete the course. However the specificity of the model in this instance is 32%. It only correctly identifies students who do not complete the course on 32% of occasions. The sensitivity of the model to predict students who partially complete the course is 26%, while the specificity in this instance is 83%. Similarly, the sensitivity of the model to predict students who drop the course is 0% and the specificity 98%. Based on these figures the ordinal model appears

to be superior to the logistic model in predicting students who complete the course. In fact the model appears to predict completing students much more effectively than those who fail to complete. This may be due, in part, to the explanatory variables available. The set used in this chapter appear to be more effective in the prediction of achievement for complete students, than for incomplete students.

4.4.3 Summary

In this section, models of educational achievement and progression were validated against a subsequent semester offering of the course TPP7181. This validation assessed the predictive accuracy of these models and more specifically their ability to generalize beyond the data in which they were estimated. Several statistics were used to assess this predictive validity and these tended to confirm earlier conclusions that were based upon measures of explained variance. For example, the CVCs for both the linear and beta models were almost equal as were their respective measures of explained variance. In one instance, this validation did provide an unexpected result. Statistics applied to the ordinal model of progression seemed to suggest that it was superior to the logistic model, this is despite the two having similar pseudo-R^2 measures.

4.5 Conclusion

In this chapter generalized linear models and other linear models, were applied to the performance data of students enrolled in a tertiary preparatory mathematics course. The models, outlined in Chapter 3, were applied to various types of performance data and for various subsets of the student group. In the first instance achievement data for students were modelled. In particular, both a standard linear regression model and a model based upon the beta distribution, were applied to the achievement data for those students who completed the course. The former model was able to explain 35% of

the variation in student achievement data, which is comparable with results obtained in similar studies (see for example the meta-analysis by Robbins et al. [84]). A Tweedie model and a Tobit model were both applied to the achievement data of those students who did not complete the course. Unfortunately there was insufficient information available to explain the variation in achievement for this group of students.

Several models were applied to the progression measures, which included whether students had completed the course and the number of assessment items that they completed. Again, there was insufficient information available to explain the variation in these data, with all models explaining less than 10% of the variation in student performance. It is likely that with larger samples, these apparently inadequate models might gain enough statistical power to be of use.

In line with evidence from the literature, cited in Section 2.2, prior knowledge was a statistically significant explanatory variable in all models of performance that were analyzed in this chapter. Other measures of prior knowledge, such as the highest level of education, were also shown to be statistically significant predictors of performance. The length of time since a student last completed formal study in mathematics seemed to positively influence performance, with more mature students being more likely to complete the course and gain higher results. This result is not uncommon in the literature (see for example Cantwell et al. [19] and Hall & Marchant [41]) although in some instances age has been shown to negatively influence performance (Pokorny & Pokorny [80]). Similarly, female students who completed the course were more likely to gain higher marks than their male counterparts, this also is confirmed by the literature (see for example Cantwell et al. [19] and De Berard et al. [28]). All conclusions derived from the application of these models, especially the models of progression data, need to reflect the poor degree of model fit (as estimated by the pseudo-R^2). In most instances there simply was not enough information to explain the variation in the response.

In this chapter the issue of violations in model assumptions and its effect on model conclusions was also discussed. In particular a linear model was fitted to the achievement data for the entire group, despite the obvious lack of Normality (see Figure 4.3). A second linear model was fitted to the achievement data for students who had completed the course (and these appeared to be closer to Normal). The estimated effect sizes of prior knowledge for these models were not significantly different. It would appear that, depending on the purpose of the statistical model, such violations of Normality may be inconsequential. In the next chapter, simulation methods will be utilized to address this issue in more detail.

Chapter 5

Simulation results

In the last chapter various linear models were applied to performance data obtained from students in a TPP mathematics course. In this chapter simulation methods are used to determine the appropriateness of standard linear regression models in an educational context and especially in situations where violations of model assumptions occur. In particular Section 5.1 examines the issue of violations in the assumption that the errors in the linear regression model are Normally distributed. This section then seeks to determine whether there is any practical benefit from using alternative linear models in preference to a standard linear regression model. Section 5.2 addresses the issue of potential bias that might occur when data from students who do not complete the course are omitted.

5.1 Violations of Normality

Linear regression is used extensively in educational research, however very few researchers report on the underlying assumptions of the model as outlined in Section 2.3.2. For example, in a sample of 14 papers published in educational and psychological journals that used standard linear regression models, only three report on the Normality of the underlying response

variable and two of these also report on the homoscedasticity (constant variance) of the data. As discussed in Section 2.3.2 the least squares estimates in linear regression have been shown to be the best linear unbiased estimates obtainable, provided that the Gauss–Markov conditions, which do not include Normality, are met. The sampling distributions of these estimates, however, are derived on the basis that the errors are Normally distributed. All of the papers in the sample mentioned above do report on the significance of the model in question (citing an F-statistic) and the significance of the coefficients (citing t or p statistics where appropriate) and have therefore relied on the Normality assumption. Of issue is whether violations in this assumption will unduly influence the results of such significance tests. Bohrnstedt & Carter [11, p. 123] cite evidence to show that when conventional regression techniques are used, the actual distribution of the error terms will have little effect on the value of the t-statistics, provided the sample is large. Further Sen & Srivastava [85, p. 106]) show that even in the absence of Normality and for large samples it is possible to derive sampling distributions for the estimates obtained from standard linear regression. It would appear that in relation to the Normality assumption, linear regression techniques are quite robust (that is insensitive to model violations), especially for large sample sizes.

In Chapter 2 it was reported that in studies of education 'extremes of asymmetry and lumpiness are more the rule than the exception' (Micceri [67, p. 161]). In fact, due to the correlations between items within an achievement test, the total test score cannot be Normal (Nunnally (1978), cited in Micceri [67]). Given the obvious violations of the Normality assumption in educational research and the proceeding discussion regarding the robustness of linear regression estimates, it is necessary to ascertain whether these violations in fact make any practical difference to the application of standard linear regression models to educational performance data.

Violations in the Normality assumption can arise from a number of sources:

- The data may be only defined for a domain which is some subset of all real numbers. This is the case with proportion data, which are only defined for the domain $0 \leq x \leq 1$.
- The distribution of the data may be skewed, that is asymmetrical.
- The distribution of the data may have a peak which is less pronounced than the Normal distribution (kurtosis).

With the widespread availability of software to fit linear models that are based on non-Normal error distributions, it is surprising that few researchers in the educational area consider their use. As discussed in Chapter 3, such models can, for example, accommodate skewed data more appropriately than merely using a standard linear regression model and thereby avoid the issues relating to Normality that were discussed above. One of the key issues of interest in this section is whether it is more appropriate to use these alternative linear models instead of simply applying standard linear regression models to non-Normal data. This appropriateness, however, depends on the original purpose of the model. In this study two commonly used purposes for regression models in the educational context are considered, namely:

1. The use of regression models to predict student achievement and in particular to identify 'at-risk' students.
2. The use of regression models to estimate the effect size.

In this simulation only one aspect of non-Normality is examined and this is the mis-specification of the error distribution. Simulated data will be generated using a skewed theoretical distribution (in this instance the beta distribution) with a known linear predictor. A standard linear regression model will then be fitted to these data. The degree of model fit between the linear predictor and the response variable in the generated data will be adjusted to see if the 'noise levels' so common in educational research

influence the appropriateness of the fitted linear model. In particular this simulation seeks to answer the following questions:

1. Is it more appropriate to use an alternative linear regression model rather than a standard linear regression model when the data are skewed?

2. To what extent does sample size influence the utility of the standard linear regression model when it is applied to skewed data?

5.1.1 Simulation methodology

As discussed in Section 3.2.3 and demonstrated in Section 4.2.1, the beta distribution is suitable for modelling educational performance data. In particular, the beta distribution can assume quite skewed forms.

The density of the beta distribution, shown in Equation (3.15) can be reparameterized using $\alpha = \mu\phi$ and $\beta = \phi(1-\mu)$. The reparameterized density will therefore be given by:

$$\text{Beta}(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}. \qquad (5.1)$$

Based on equations (3.16) and (3.17), it can be shown that the expected value and variance of a beta distributed random variable are:

$$E(Y) = \mu$$

and

$$\text{var}(Y) = \frac{V(\mu)}{1+\phi}$$

respectively, where the function $V(\mu) = \mu(1-\mu)$ can be regarded as the variance function for the distribution, and the parameter ϕ a measure of

dispersion. Ferrari & Cribari-Neto [38] term ϕ a precision parameter, as the greater the value of ϕ the smaller the variance.

In order to assess the influence of non-Normality on standard linear regression models, populations of variables (X,Y) were generated so that $Y \sim \text{Beta}(\mu, \phi)$ and $\text{logit}(\mu_i) = \beta_0 + \beta_1 x_i$. The parameters were adjusted so that the degree of fit in the linear model and the skewness of the underlying response distribution could both be altered. Linear models that are applied to skewed distributions are unlikely to have error distributions that are normal, especially at either end of the explanatory variable domain. The degree of fit was measured using the pseudo-R^2 as defined in Equation (3.11).

Using this alternative parameterization of the beta distribution, the skewness of the beta distribution, as shown in Equation (3.18), can be rewritten as:

$$\gamma_1 = \frac{2(1-2\mu)}{\phi+2}\sqrt{\frac{\phi+1}{V(\mu)}}. \tag{5.2}$$

Consequently the skewness is dependent on both the precision parameter ϕ and the mean of the response μ. The degree of fit of the model can be altered by varying the parameter ϕ, however in doing this the skewness will also vary. This fact has influenced the specific details of the methodology, which are detailed below (R-code is shown in Appendix A).

1. A beta model was fitted to the performance data used in Chapter 4 for complete students, using only M-test as a regressor. From this, estimates for the constant b_0, effect b_1, and precision parameters ϕ_0 were obtained for the model:

$$\text{logit(Score)} = b_0 + b_1 \times \text{M-test}.$$

2. Using b_0 and b_1 as starting points, parameter values were chosen for the constant β_0 and the effect β_1, of the linear predictor in the proposed

simulated population. These values were chosen specifically so that the final distribution of performance values was skewed.

3. A random sample of N values from the set of M-test scores (m) in Chapter 4 was obtained. Using the parameter values above, a vector of linear predictor values $\eta = \beta_0 + \beta_1 m$ was then created.

4. Using Equation (5.1) a vector of N random performance scores y was generated from a beta distribution, with a mean given by:

$$\mu = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

and precision parameter ϕ. The later was some multiple of the initial estimate ϕ_0 and was adjusted in order to create linear models with an appropriate pseudo-R^2 value.

5. In this way it was possible to create a set of N points (m_i, y_i) to which a standard linear regression model could then be applied.

In order to evaluate the linear model in this instance, a simple measure was developed. A 95% confidence interval for the mean conditioned on each value of the explanatory variable $(\mu_i|m_i)$ was calculated for the linear model. The true value of the mean μ_i was then compared with this confidence interval for each value of m_i. The number of instances when the confidence interval 'captured' the true mean was expressed as a proportion of the total number of m_i and this was termed the 'effectiveness' of the linear model.

The above procedure was repeated 100 times for different parameter values and sample sizes. The results of this simulation are reported in the next section.

5.1.2 Results of simulation

The mean effectiveness of the linear model, as defined above, is shown in Table 5.1 for situations where the skewness of the response variable was low ($\gamma_1 \approx -1$), moderate ($\gamma_1 \approx -1.5$) and high ($\gamma_1 \approx -2$) and for low model fit (pseudo-$R^2 \leq 35\%$). As can be seen, the degree of skewness in the performance distribution will influence the extent to which the linear model is able to capture the true means. As expected, the more skewed the distribution the lower the effectiveness of the linear model. For example, in a situation of high skewness and large sample size ($N = 500$), the linear model is effective on 36% of occasions, while in a situation of low skewness and large sample size it is effective on 71% of occasions. The results also show that in smaller sample sizes the linear model appears to be more effective. This later result, although seemingly counter-intuitive, is a result of the decreased power that comes from smaller samples, and so is not unexpected. In such instances of small sample size and high variability, there is insufficient statistical power to differentiate between the models. The top plot in Figure 5.1 demonstrates the application of the linear model to a highly skewed performance distribution. In this plot it can be seen that the 95% confidence interval for the mean (as determined from the linear model) rarely captures the actual mean (as shown by the unbroken curve). Moreover such capture occurs primarily in the 'middle ranges' of the explanatory variable. The bottom plot in this figure demonstrates the application of a linear model to low skewed performance data. As can be seen from this later plot, the effectiveness of the linear model is improved in applications where the skewness of the performance distribution is low.

5.1.3 Discussion

In this section non-Normal data were generated using a beta distribution. Through a simulation process it was possible to create sets of data (m_i, y_i)

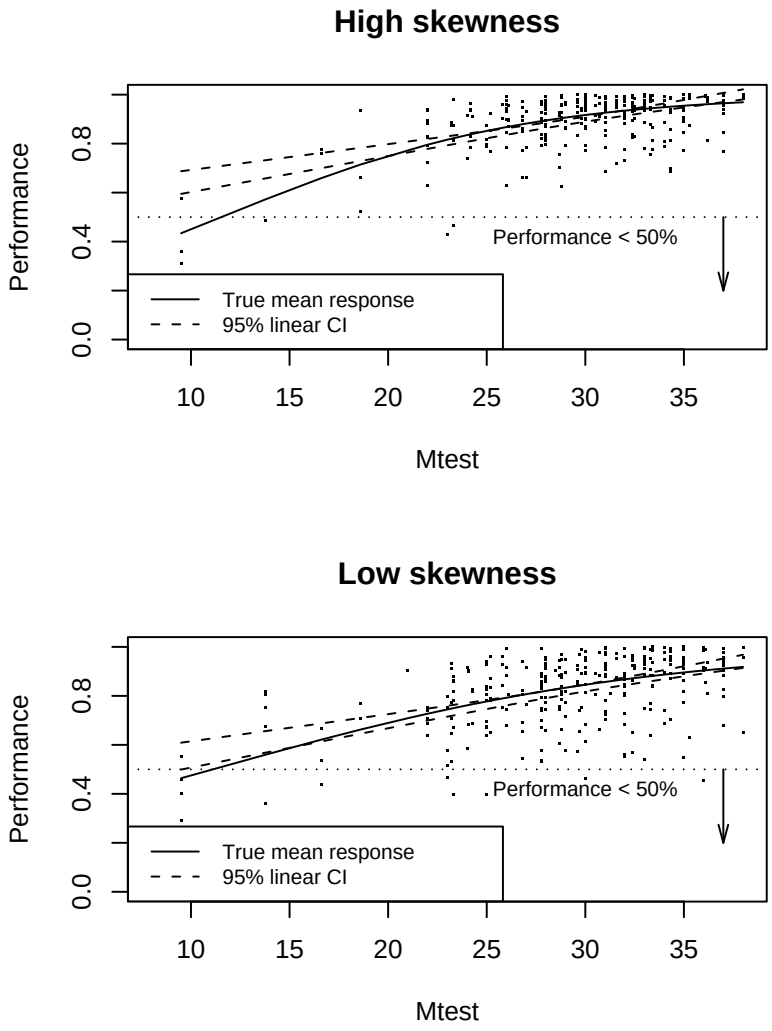

Figure 5.1: Application of the linear model to both high and low skewed distributions. The 95% confidence interval is shown for the linear model as dashed lines and the true (generating) model as an unbroken line.

Situation	Sample size	Mean effectiveness
Low skewness	100	89%
Moderate skewness	100	81%
High skewness	100	70%
Low skewness	300	77%
Moderate skewness	300	60%
High skewness	300	54%
Low skewness	500	71%
Moderate skewness	500	46%
High skewness	500	36%

Table 5.1: Effectiveness of the linear model for different degrees of skewness and sample size

for which y_i was sampled from beta distributions of varying skewness. It was also possible to vary the strength of the linear relationship between m_i and logit(y_i). A linear model was then fitted to these data and its effectiveness, as defined earlier, was calculated.

As can be seen from the upper plot in Figure 5.1, when the overall performance distribution is highly skewed the distribution of observations conditioned on any given value of the explanatory variable $y_i|m_i$, will not be uniform and will in fact be highly skewed at each end of the distribution. Consequently the assumption in standard linear regression that the errors are Normally and identically distributed will be violated. In such instances of extreme skewness, the 95% confidence interval for the mean performance conditioned on values of the M-test, often failed to capture the true mean. In other words such a violation of the Normality assumption appears to lead to a reduction in the utility of the linear model. This problem is magnified at the extreme ends of the explanatory variable domain, where quite large

discrepancies between the true value and the 95% confidence boundaries can be observed. Arguably in educational research it is these areas that are of particular concern. One application of this is in the use of pre-tests, such as the M-test, to predict students who are at risk of failing the course (achieving a performance score less than 50%). It can be seen from both plots in Figure 5.1 that the linear model fails to identify any 'at-risk' students, while the true model identifies students who score approximately less than 12 in the M-test to be at-risk of failing the course.

Statistical models in the educational area are commonly used to estimate the effect size of the explanatory variable in question. A linear model implies a constant effect size, which is often untenable in the educational context. For example, a change in M-test results from zero to five marks out of a total of 38, is unlikely to produce any noticeable difference in performance. A change in marks from 20 to 25, however, may be more likely to produce a change in performance. It can be seen from Figure 5.1 that the effect sizes (as determined by the gradients of both the linear and true models) are very similar in the lower plot (situation with low skewness) and less similar in the upper plot. Again major differences occur in the tails of the distribution. For example, it can be seen from the top plot of Figure 5.1 that an increase in M-test results of 5 marks (from 10 to 15) will result in an 18% improvement in predicted performance. In this instance, however, the linear model only predicts an increase in performance of 5%.

The above discussion neglects another important aspect that relates to statistical modelling, that is the interpretability of the model. For small enough values of m_i the linear model may predict negative performance and similarly for high enough values it may predict performance greater than 100%. It would seem that in applications related to prediction or estimation of effect size, the linear model is not appropriate for performance data at either end of the range of the explanatory variable, and this problem is exacerbated in highly skewed distributions of the response variable.

5.2 Issues relating to incomplete data

The other key issue of interest in this study regards the treatment of performance data for students who fail to complete the course. In the current context, such a group can constitute over one half of the entire student group. Most studies in the literature do not indicate how they treat such data as presumably they form a smaller proportion of the total student group. For example, in a sample of 32 studies dealing with the educational performance of tertiary students: 18 failed to make any mention of how they dealt with the performance data for incomplete students; six restricted the scope of their study to the performance of completing students; two dichotomized the data so that presumably incomplete data were included in one of the categories; and the remainder treated the incomplete data as missing. In the later category, Cantwell et al. [19] ignored the data from 500 students (approximately 6% of the total) as 'many of them would not have remained at University beyond a week or two' [19, p. 223]. In another instance, Lane et al. [55] used results from only 65 of a total of 137 students who had agreed to participate in the study, presumably from discarding observations that contained missing values. Such practices may be the most appropriate way to deal with this data, especially if it is the intention to generalize the results to include only complete students. At issue in this section, is whether the data for incomplete students should be included in models of achievement or treated separately. Bosshardt [12, p. 112] cites evidence from Becker & Powers (2000) who found that ignoring such data could substantially alter the effect of class size on learning (from no effect when the data are ignored to a negative effect when they are included).

In Sections 4.2.1 and 4.2.2 achievement for complete and incomplete students was modelled separately. It was found that, although the M-test was a statistically significant predictor in both models, it was more weakly predictive for incomplete students. Moreover, factors that predicted achievement for complete students did not predict achievement for incomplete students.

These results suggest that it is appropriate to model the two groups separately. Nevertheless, results from Section 4.2.3 demonstrated that there was in fact no statistically significant difference between the effect size of M-test on achievement for complete students and the effect size of M-test on achievement for a model based on the entire group of students.

In this section simulation is used to investigate the most appropriate statistical methods for dealing with incomplete data. In particular two issues are investigated:

1. Whether it is appropriate to include the data of incomplete students and apply a linear model to them, rather than to ignore the data for these students (reported below); and,

2. If the data for incomplete students is to be treated separately, the most appropriate way to deal with the data for the students who drop the course and who are notionally given an achievement result of zero (reported in Section 5.2.2).

5.2.1 Incomplete data in models of performance

As discussed in Section 4.2.3 the performance distribution for the entire group in the current context can be regarded as a mixture of performance results generated from two specific sub-groups. The performance of students who failed to complete the course (termed 'incompletes') can be modelled using the Tweedie distribution (see Section 4.2.2) and the performance of students who completed the course (termed 'completes') by either the beta distribution or the Normal distribution (see Section 4.2.1). This treatment of the data forms the basis of the simulation in this section. More specifically:

1. A simulated sub-population of incomplete performance data ($N = 1000$) was generated from a Tweedie regression model, using parameters that were estimated from the performance data of incomplete students in Chapter 4.

2. A simulated sub-population of complete performance data ($N = 1000$) was generated from a standard linear regression model using parameters that were also estimated using the performance data of complete students in Chapter 4 . These data was truncated at 100%, the upper limit of the performance scale (when expressed as a percentage).

3. Different mixes of these two data-sets were combined in order to obtain a population of 1000 observations.

4. A linear model was then applied to this mixture and compared with the linear model that generated the performance results of complete students.

Results of the simulation

Simulated incomplete performance data were generated using a Tweedie regression model with parameters: $p = 1.08, \phi = 6.3$ and a linear predictor $\eta = 0.7 + 0.06m$ (where m is the score in the M-test). The linear model in this instance had a pseudo-R^2 value of 0.086. The distribution of performance values for this sub-population is shown as the top graph in Figure 5.2. Similarly, complete performance data were generated using a standard linear regression model with a variance of $\sigma^2 = 172$ and a linear predictor $\eta = 33 + 1.3m$ ($R^2 = 0.22$). Simulated performance values in excess of 100 % were discarded and further sampling occurred until all values were within the appropriate range. The distribution of this sub-population is shown as the middle graph in Figure 5.2. Initially these two sub-populations were sampled so that 50% of observations were incomplete and 50% complete (this was close to the ratio obtained from the existing data used in Chapter 4). The distribution of this mixture is shown as the lower graph in Figure 5.2 and is similar in appearance to the distribution of results shown in Figure 2.2.

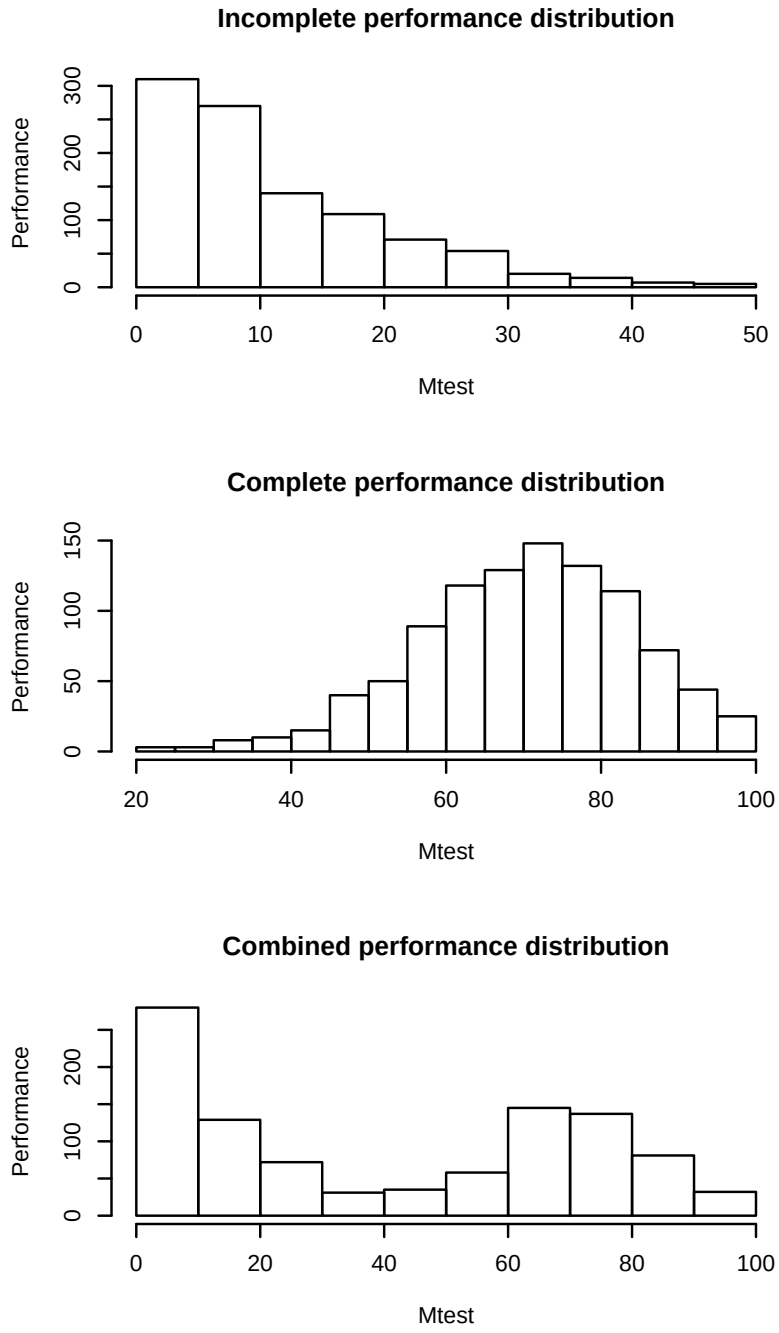

Figure 5.2: Distributions of sub-populations (in the ratio 1:1) and combined population $N = 1000$

Four different 'mixes' of sub-populations were then created, these contained: 20%, 40%, 60% and 80% of incomplete performance data respectively (shown as Graphs 1 to 4 respectively of Figure 5.3). In each instance a linear model was fitted to the entire distribution and this was compared with the linear models used to create the two sub-populations. Only in Graph 1 of this figure does the 95% confidence interval for the gradient of the overall model capture the gradient of the model that generated the data for the complete students. In other words, for a small proportion of incomplete data, there appears to be little difference in the two models. To further investigate this, population mixes that ranged from 1% of incomplete data to 99% of incomplete data were generated. As before, a linear model was applied to the mix and then compared with the linear model used to generate the complete data. In most instances up to a mix of 50% incomplete, the 95% confidence interval for the gradient of the overall model captured the gradient of the model that generated the complete data. So apparently, there is little practical difference in omitting the data for incomplete students or retaining these data, in mixes with up to 50% incomplete.

Discussion

In this section distributions of performance data that included various mixes of complete and incomplete performance were created. At issue, was whether the parameters of the commonly used linear model would be affected by the omission of data from incomplete students. The results indicate that unless there was a large proportion of incomplete data in the overall mix, effect sizes for models applied to all data and models applied to complete data, would not differ substantially. The large changes in the intercept term when such data were excluded, meant that models of prediction would be more adversely affected.

The results of this simulation, however, are inconclusive. It is likely that with the high variance encountered in this simulation (and indeed in most

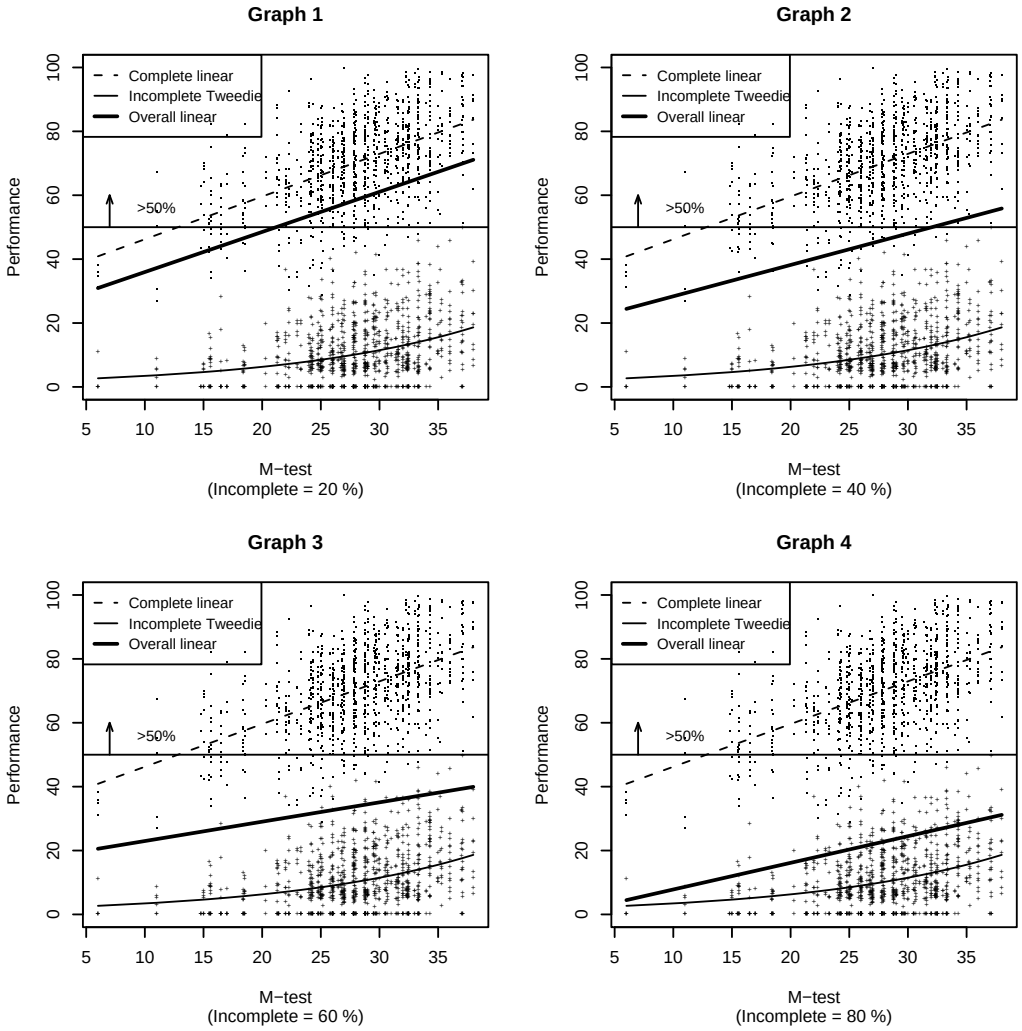

Figure 5.3: Comparison of models applied to sub-populations and to a mix of different proportions of these sub-populations. Graph 1 shows the overall linear model when 20% of incomplete data is included, Graph 2 when 40% is included, Graph 3 when 60% is included and Graph 4 when 80% is included.

educational contexts), there is insufficient statistical power to distinguish between any two models. Given this lack of power, other measures of model appropriateness may need to be considered. For instance, the assumption of the linear model that the effect is constant, may not be appropriate. Indeed, a model based on the beta distribution may be more appropriate (see discussion in Section 5.1). Similarly, as mentioned earlier, it might not be valid to even combine these data-sets. The simulated achievement data for incomplete students was only weakly explained by their M-test results. Their achievement data reflect the length of time that they remained in the course. This in turn may be explained by factors that do not necessarily explain achievement. So while the omission of data for incomplete students may or may not influence overall model parameter estimates, there are probably good reasons for treating the two sets of achievement data separately. The next section addresses this issue further, with an analysis of data obtained from incomplete students.

5.2.2 The problem of exact zeros

In the previous section, data for complete and incomplete students were simulated separately and then combined. In this section, the modelling of performance for those students who fail to complete the course is discussed and in particular the influence of the exact zeros obtained by those students who withdraw without completing any assessment item (drop-outs). As was mentioned in the previous section, the presence of exact zeros in many statistical models is problematic. For example, the beta model is not defined for zero values. Similarly, the presence of a large number of exact zeros in a standard linear regression model will violate the Normality assumption. In this section, the issues related to the exclusion or otherwise of exact zeros are examined.

In section 4.2.2, a Tweedie regression model was applied to the educational performance data of incomplete students. The model was only able

to explain approximately 9% of the variation in performance scores and so could not be reliably developed. If it can be assumed that it may be possible to find explanatory variables that better account for the variation in performance of these students, then a simulation based upon the Tweedie model may be useful for the task at hand.

In this simulation a Tweedie regression model was developed so that it would explain in excess of 20% of the variation in student performance scores. This model was then used to generate data that could be analyzed using other methods. More specifically, in this simulation:

1. 1000 explanatory values m_i were sampled with replacement from the set of M-test scores used in Chapter 4.

2. A vector of linear predictors was calculated, with parameters chosen in order to achieve an increased model fit statistic.

3. A Tweedie distribution was then used to generate suitable response variables y_i.

4. Simulated observations that contained exact zeros were then excluded and alternative models were then fitted to the remaining data.

Results of the simulation

In this instance, simulated incomplete achievement data were generated using a Tweedie regression model with parameters: $p = 1.09$, $\phi = 6.3$, and $\mu_i = \exp(\eta_i)$. The parameters in the linear predictor $\eta_i = -0.8 + 0.11m_i$ were chosen in order to achieve a pseudo-R^2 of 23.5%. In this way it was possible to generate data (m_i, y_i) that were weakly related and that included a large proportion of exact zeros (26%). The distribution of these simulated achievement scores is shown in Figure 5.4. All observations that contained exact zeros were excluded from this data-set and a linear model was then applied to this subset. As the beta distribution can accommodate quite skewed

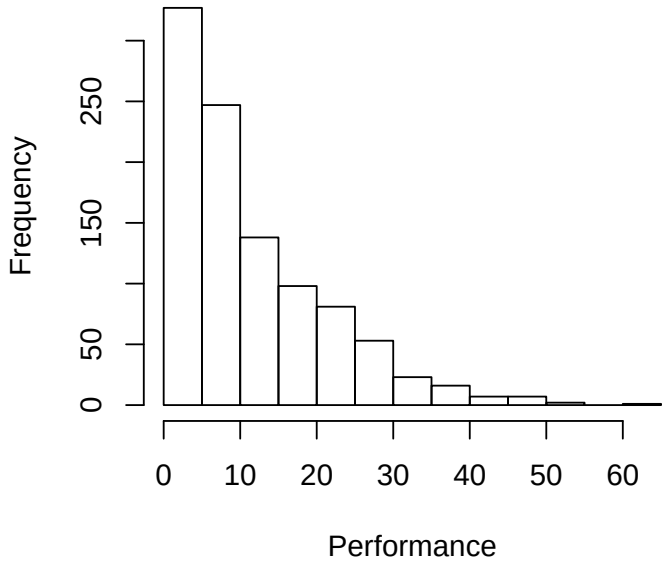

Figure 5.4: Distribution of simulated achievement scores for incomplete students

data, such as these, a beta linear model was also applied to this subset. A plot of these models together with the generating Tweedie model is shown in Figure 5.5. The top plot (Graph 1) in Figure 5.5 shows the predicted mean response from the linear model compared with the true mean response as generated by the Tweedie model. It can be seen that the linear model is inappropriate, as it predicts negative performance values when M-test results fall below 10. Moreover, it assumes a constant change in achievement over the entire domain of the explanatory variable m_i, an assumption that is obviously not appropriate in this instance. The lower plot (Graph 2) of Figure 5.5 shows the predicted mean response for the beta model compared with the true mean response. This model is far more appropriate than the linear model for this subset, as it does not predict negative achievement. The beta model, however, does predict a more constant change in achievement than is suggested by the generating model. For example, the Tweedie model predicts a change in achievement of approximately eight percentage points when M-test results rise from 35 to 38, while the beta model only predicts an increase of three percentage points.

5.2.3 Discussion

The Tweedie model is able to generate data that include exact zeros and simulate distributions that might appear in this particular context. In fact it was demonstrated in Chapter 4 that such a model could be applied to the performance data of students who fail to complete the course. Unfortunately in that instance, the strength of association between the explanatory variables and the response variable was low. For this reason, it was felt that the application of the Tweedie model to educational performance data, although novel, was merely academic. In this simulation it is assumed that it may be possible to obtain stronger predictors of performance for these particular students. Given this assumption, it can be seen that the common practice of excluding exact zeros from the analysis of such performance data will clearly

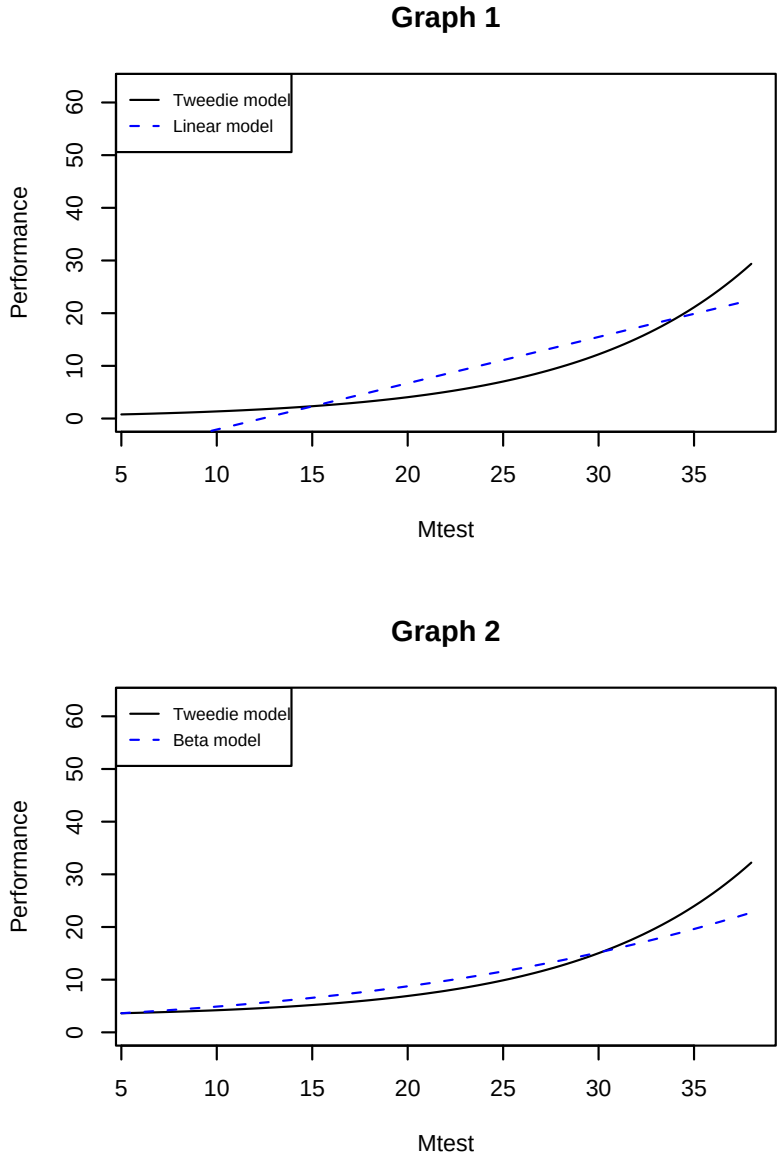

Figure 5.5: Models applied to non-zero achievement data generated by a Tweedie regression model. Graph 1 shows the standard linear regression model applied to all non-zero data and Graph 2, a beta model applied to the same subset.

produce biased results. Whether such bias is of practical concern, though, is another issue. The Tweedie model predicts that a student who scores 5 in the M-test will have a performance of almost 1%, while the beta model predicts a performance result of almost 3%. Both results are very low and it is questionable whether any educational assessment instrument could measure performance to this degree of fidelity. This simulation also demonstrates that it is inappropriate to apply a linear model to these data.

5.3 Conclusion

In this chapter simulation methods were employed in order to assess the appropriateness of the standard linear regression model, in situations of non-Normality and situations where there were a high proportion of incomplete students. In the first section of the chapter a beta regression model was used to generate sets of points (m_i, y_i). These points were generated in such a way that both the strength of fit between m_i and y_i, and the skewness of y_i, could be varied. In each instance a linear model was fitted to the data and then compared with the beta model that generated the data. In the second section of this chapter the issue of missing data was addressed, and in particular data associated with students who fail to complete a course. The simulation results were inconclusive, however it was felt that such data should be modelled separately. The issue of how to deal with the data from students who withdraw from the course without any performance measure, was also discussed. It was found that excluding such data from models based on the performance of incomplete students, would produce biased model results. It was unlikely, however, that such bias would be of any practical significance.

Standard linear regression models are widely used in the educational research area and as stated earlier, few researchers report on the underlying assumptions for these models. Many report the significance of regression estimates without reporting the Normality or otherwise of the data, despite

these measures being based on distributional assumptions. Educational performance data are seldom Normal. They are usually measured on a percentage scale, in other words they contains upper and lower bounds. In many instances by the very way they are constructed, that is as a composite of smaller test marks, educational data cannot be Normal. One of the motivations of this study was to determine whether such departures from Normality and the failure of the research community to report these, are of concern. Models based upon educational performance data are usually 'noisy'. That is they are typically characterized by pseudo-R^2 values in the order of 35%. In such instances the accuracy or precision of any model is compromised and issues such as the effects of non-Normality on estimates become clouded. In this simulation only one aspect of non-Normality was considered, that being the skewness of the response variable. It was found that in general, the performance of a standard linear regression model, whether for predictive purposes or for the establishment of effect size, was adequate in the mid-domain of the explanatory variable, irrespective of the degree of skewness in the response variable. It was in the tails of the distribution, however, where large discrepancies occurred. In many instances educational research is specifically interested in these domains of the explanatory variable, and the one mentioned in this chapter was the use of pre-tests to identify at-risk students. When a beta model was applied to skewed data, rather than the linear model, there was a much greater likelihood that the former would correctly identify such students. Whether such identification is accurate, is another issue and is reported in Carmichael, Dunn & Taylor [20].

The issue of how best to treat missing data is important in any statistical models. In many instances extrapolation methods can be used to infer the value of missing data in explanatory variables. For example Smyth [88] successfully used a two stage algorithm called the Expectation/Maximization (EM) algorithm to estimate missing values of the explanatory variables in a large study of student performance data. In the current context the issue is

different, in that the missing values have occurred in the response variable. Moreover, the performance scores of incomplete students may reflect more the length of time they spend in the course than their actual performance in mathematics. In this chapter, only the statistical implications of including the data of these students in models of educational performance, was considered. It was found that the actual purpose of the model would influence the treatment of such data. If a model is used for predictive purposes, for example to identify at-risk students, then it would be foolish to exclude the data for incomplete students. If on the other hand, the model is to be used to examine effect sizes, then unless the proportion of incomplete students is greater than 30%, there is little point in including such data. Indeed, there may be valid reasons for excluding such data, as the factors that cause such students to drop the course and have low performance scores may be quite different to the factors that explain performance for the mainstream (complete) student.

The issue of how best to treat the performance of students who fail to produce any measure is probably academic. Denoting such performance as a zero would seem a reasonable thing to do, however this produces problems with many statistical models. This is because statistical models are based on distributions that either do not contain large numbers of zero (for example the Normal distribution) or are simply not defined for zero (for example the beta distribution). The Tweedie regression model can accommodate such situations. This model may provide a novel and useful method for dealing with educational performance data that include exact zeros. Its utility, however, relies on the existence of reasonably strong explanatory variables, and these were not available in this study. Consequently results obtained from using such a model are unlikely to be of any practical benefit in the current context.

In this chapter, simulation approaches were used to assess the appropriateness of the standard linear regression model, when it is applied to educa-

tional performance data. In the main, the measures of appropriateness were based on a comparison of model parameter estimates. In the educational context, however, the inherently high variation creates a situation where there is insufficient statistical power to distinguish between models. Consequently other measures of model appropriateness need to be considered. For example, is it appropriate for the model to assume that the performance measure changes constantly with corresponding changes in the explanatory variable? Does the model allow for predicted values outside of the accepted range of performance scores (that is greater than, say 100%)? Based on such measures of appropriateness, the results from this chapter indicate that in many instances, it is not appropriate to use the standard linear regression model.

Chapter 6

Summary and discussion

6.1 Summary

The purpose of this study, as outlined in Chapter 1 was to examine the widespread practice in educational research of using statistical models to explain the variation in student performance data and in particular to evaluate the application of such models when the unique properties of educational performance data are considered.

In Chapter 2 of this dissertation, a review of the literature as it relates to the modelling of educational performance data was undertaken. The first section of this chapter examined the problems associated with actually defining and then measuring educational performance data. It was noted that at the course level such data could include either measures of student achievement or measures of student progression. The second section of this chapter then sought to identify possible predictors of student performance. The literature provides countless examples of such predictors, and it was noted that for convenience these could be classified according to a framework established by Snow et al. [91] who identify three major categories of such predictors, namely cognitive, conative and affective. The third section of this chapter examined common methods for modelling educational performance data. It

was found that the use of complex models such as those encountered in structural equation modelling, was commonplace in this area of research. These models attempt to explain the complexity that undoubtedly exists in this context. Due to their complexity, however, such models were often found to be difficult to interpret. Moreover such models in most cases require distributional assumptions that are not easy to satisfy. The use of simple models in educational research was then discussed. It was noted that in instances where the researcher seeks to ascertain the influence of only a few factors on performance, such models would be preferable to the complex ones discussed earlier in the chapter. The fourth section of this chapter then identified the particular context in which this study was based. The unique aspects of the performance data of students studying preparatory mathematics were identified, including the large number of students who fail to complete the course. Finally a justification for the methodology used in this study was provided.

In Chapter 3 of this dissertation a review of simple statistical models was undertaken with the purpose of establishing and justifying the methodology used in this study. In the first section of the chapter the framework associated with the theory of generalized linear models was discussed. It was noted that alternative models within this framework may be more suitable for educational performance data than the standard linear regression model (also a generalized linear model). The second section of this chapter examined statistical models that may be suitable for modelling the non-Normal data that typify educational performance data. It was noted that, for example, a linear model based upon the beta distribution could model skewed data and data that contained upper and lower bounds, as in achievement data that are measured on the percentage scale. The third section of this chapter then discussed possible methods for modelling the exact zeros that can be obtained in educational performance data and that typify the particular data in this context. This section explored two models that have successfully been used to model such data in non-educational contexts, these were the Tweedie

regression model and the Tobit model. The final section of the chapter then addressed the third project objective, and in particular outlined how simulation methods could be used to ascertain the appropriateness of the standard linear regression model when certain model assumptions are violated.

In Chapter 4 of this dissertation the results of the application of statistical models to educational performance data were presented. In the first section of this chapter, the data-set used was introduced. This data-set included the performance data obtained from a group of students studying a preparatory mathematics course at the University of Southern Queensland. It also included a number of possible explanatory variables, such as their performance in a course pre-test and other demographic details. In the second section of this chapter, statistical models were applied to the achievement data of these students. In particular different linear models were applied to the achievement of the whole group and then to different sub-groups, as defined in Section 3.4. It was found that such linear models seemed to perform well (they were comparable with studies reported in the literature) when they were applied only to the achievement of students who completed the course.

In the third section of Chapter 4, linear models were then applied to progression data for these students. Both models used and reported in this section had pseudo-R^2 values far less then similar studies reported in the literature where R^2 values usually exceed 20%. It would appear that there were not adequate measures available to explain these aspects of student performance. This issue will be discussed further in the next section of this chapter. The fourth section of this chapter reported the results when the models used earlier in the chapter were applied to an independent data-set. This validation of the models was able to reinforce the point that there were insufficient measures available to predict the performance of low achieving students.

In Chapter 5, simulation methods were used to answer the third project objective, as outlined in Section 1.2. The first section of this chapter exam-

ined the appropriateness of applying the simple linear regression model to non-Normal educational performance data. Simulation methods were able to show that it is more appropriate to use alternative linear models in situations where the data was highly skewed. More importantly, large discrepancies occurred between the standard linear regression model and other models in the 'tails' of the distribution, so that any problems with applying the standard linear regression model to skewed data are exacerbated if the researcher is particular interested in the situation at either end of the explanatory variable domain.

The second section of this chapter examined the practice of discarding data for students who fail to complete the course. It was found that, unless there was a large proportion of such students, there was little difference between parameter estimates of models that included the data and those that did not. Such differences, however, are difficult to discern when there is high error variance in the models (as was the case). Consequently other factors need to guide the practitioner regarding their choice of regression model. It was felt, that in this instance, it would be more appropriate to model the achievement of both groups of students separately.

The last section of Chapter 5 examined the occurrence of exact zeros in educational performance data. It was found that the Tweedie regression model could provide an accurate and sensible method for estimating parameter estimates in the data for incomplete students, which contained exact zeros. The Tweedie model was much more appropriate than the standard linear regression model, which predicted negative performance data for some values of the explanatory variable.

6.2 Discussion

In the previous section, a brief overview of this dissertation was presented. In this section the findings of this study are discussed. The discussion in the

first part of this section relates to the specific context of this study, and the discussion in the second part, to the modelling of educational performance data in general.

6.2.1 Explaining performance in a TPP context

In Chapter 4 several models of performance were reported. These were based on measures of achievement and progression for students enrolled in a tertiary preparatory mathematics course. As was mentioned in this chapter, such students possessed unique characteristics that may limit the ability to generalize model results. For example, 20% were serving in correctional centres, the students were older than typical undergraduate students and a high proportion failed to complete the course, many of these dropping without completing any measure of achievement. Despite these apparent differences, however, the predictors of performance for students who completed the course were fairly consistent with those reported in the literature. For example prior knowledge, as measured using a course pre-test, was the best predictor of achievement. Similarly female students were more likely to achieve higher marks than their male counterparts.

One of the unique aspects of this data-set was the high proportion of students who did not complete the course. Several models were applied to the student performance data (both achievement and progression) of these students. In particular, two novel approaches were applied to these data the: Tobit model, which is used extensively in the econometric field; and the Tweedie model. No studies noted have attempted to apply either of these models to educational performance data. Both models were able to explain some of the variation in student performance, although the pseudo-R^2 in both instances was low. Unfortunately there were insufficient measures available in this study to explain the variation in the performance of students who fail to complete this course. This would be an important area for future research.

6.2.2 Modelling educational performance in general

The modelling of performance data is commonplace in educational research. In most cases, researchers seek to identify factors that predict student performance, in some instances these are used specifically to identify at-risk students. Techniques for achieving these aims are many and varied, but can be broadly categorized into those that utilize structural equation models and those that utilize linear regression models. The later are used extensively in educational research. For example, in a sample of 40 studies that sought to explain student educational performance, almost two thirds used some type of linear regression model as their primary methodological tool, with one half of these using a standard linear regression model. The ease of application and interpretation of standard linear regression models could explain their prevalence. Few researchers report on the underlying assumptions of the models that they have used. It is likely that researchers consider the standard linear regression model to be suitably robust, as Bohrnstedt & Carter [11, p. 142] report in 1971 'there is ample evidence to suggest that regression analysis is adequately robust except in the presence of measurement and specification error'.

The performance of students is often difficult to measure. As discussed in Section 2.1 the commonly used Grade Point Average (GPA) is problematic and measures of student achievement often rely on assessment instruments that do not have the required fidelity. Educational performance data will also contain missing or incomplete values (results from students who fail to complete their study) which may contribute to biased model estimates. Further, educational achievement, which is often the measure of performance used, is not Normal (see for example Micceri [67]). These features of the performance variable itself, undoubtedly create problems with any model of prediction. The major problem with such models, however, is the lack of identifiable and measurable explanatory variables. This dearth of predictor variables ensures that models of educational performance are characterized

by low pseudo-R^2 values (see for example Robbins et al. [84]). Moreover many explanatory variables reported in the literature are measured with error which will have a considerable influence on final model estimates (see Bohrnstedt & Carter [11]). Arguably, such problems are exacerbated in the tertiary context, where evidence (cited in Petrides et al. [79, p. 240]) suggests that the strength of the correlation between some predictors of performance (for example prior knowledge) and performance will decrease with age.

It could be argued that the low pseudo-R^2 values typically found in models of educational performance allow researchers scope to ignore inherent problems with the data. As mentioned earlier, in many cases there is insufficient statistical power to distinguish between any models of performance. Certainly, the simulation studies in this dissertation demonstrated that the influence of, for example, non-Normality on model estimates was clouded by such 'noise' in the model. Nevertheless it is argued that despite problems with these data, there is still scope for improvement in the way that educational performance is modelled. For example, in Chapter 5 it was demonstrated that the application of a standard linear regression model to achievement data that were skewed was highly problematic at either end of the performance range. This was especially the case when the model was used for predictive purposes such as to identify at-risk students. The application of a beta model to such data was shown to be more appropriate. Similarly, it was demonstrated that the Tweedie regression model had the potential to better model data that contained the exact zeros obtained when students dropped a course without producing any performance measure. No studies noted in the literature have used either of these models in the educational context.

6.3 Future research

In the last section the findings of this study were discussed, and in particular it was noted that the standard linear regression model was inadequate for

the prediction of the performance of students at either end of a skewed performance distribution. This has important implications for the widespread practice in education, of using pre-course tests to predict students at risk of failing. It was also noted in the last section that there were insufficient explanatory variables available to adequately explain the performance of students who fail to complete the course. This lack of explanatory measures also has important consequences in education, but more so in the current context, where a high proportion of students fail to complete the course. In this section, both of these issues are discussed and targeted as important areas for future research.

6.3.1 Identification of at-risk students

Many tertiary level courses require students to complete a pre-course test. In many cases, such tests are used to identify students who should be targeted for additional support. For example, in a 2001 study of mathematics, engineering and physical sciences departments in the United Kingdom [57], 68% of respondents indicated that their department used a diagnostic mathematics test in the first few weeks of semester. In the majority of these cases (70%), such tests were used to determine the level of support in mathematics required by these students. Anecdotal evidence suggests that in some instances, such tests are even used to exclude students from taking the course. Given the prevalence of such tests and the importance of ensuring that scarce learning support resources are directed to the appropriate students (or indeed that correct students are excluded), it is important that such tests are in fact able to identify at-risk students. In this study, the models of student performance presented were only able to identify at-risk students on some occasions (see discussion in Section 4.4 and also reported in Carmichael, Dunn & Taylor [20]).

It is likely that any course with a similar assessment schedule as the one reported in this study (see Table 4.2) will produce achievement results which

are skewed. Arguably assignments based on fewer learning objectives will be more skewed than an exam that is based on a greater number of such objectives. There is much less spread in smaller instruments as with fewer learning objectives to assess, there may be a tendency for students to either get them right or wrong. Moreover with less material to learn it is likely that more students will succeed in these smaller instruments (a binomial distribution with a high probability of success and a small number of trials is skewed). This was certainly the case in the current context. Figure 6.1 shows the score distributions for each of the assessment instruments used to produce the final mark and hence grade. It can be seen that in most cases the assignments, which assessed fewer learning objectives, were left skewed. On the other hand the examination, which assessed the entire course's learning objectives, was closer to Normal. The distribution of the weighted sum of such scores will likely be skewed. Given the importance of correctly identifying at-risk students, future research is needed beyond the current context, in order to ascertain whether alternative linear models can provide improved prediction validity. Of course such research would need to accompany further research into the identification of factors that lead to student incompletion. Based on the findings in Chapter 5 it would seem that beta linear models may be a possible alternative model to the standard linear regression model. Also, given the reasonably high sensitivity scores of the ordinal model (see Section 4.4), this linear model may also prove to be a suitable alternative to the standard linear regression model (although Taylor, West & Aiken [94] caution that there is substantial loss in power for ordinal linear models when the distribution is skewed and the number of categories is less than 5).

6.3.2 Dealing with incomplete performance data

One of the aims of this project, was to ascertain the most appropriate method for dealing with the performance data of students who fail to complete a course. Some evidence from the literature (see for example Bosshardt [12,

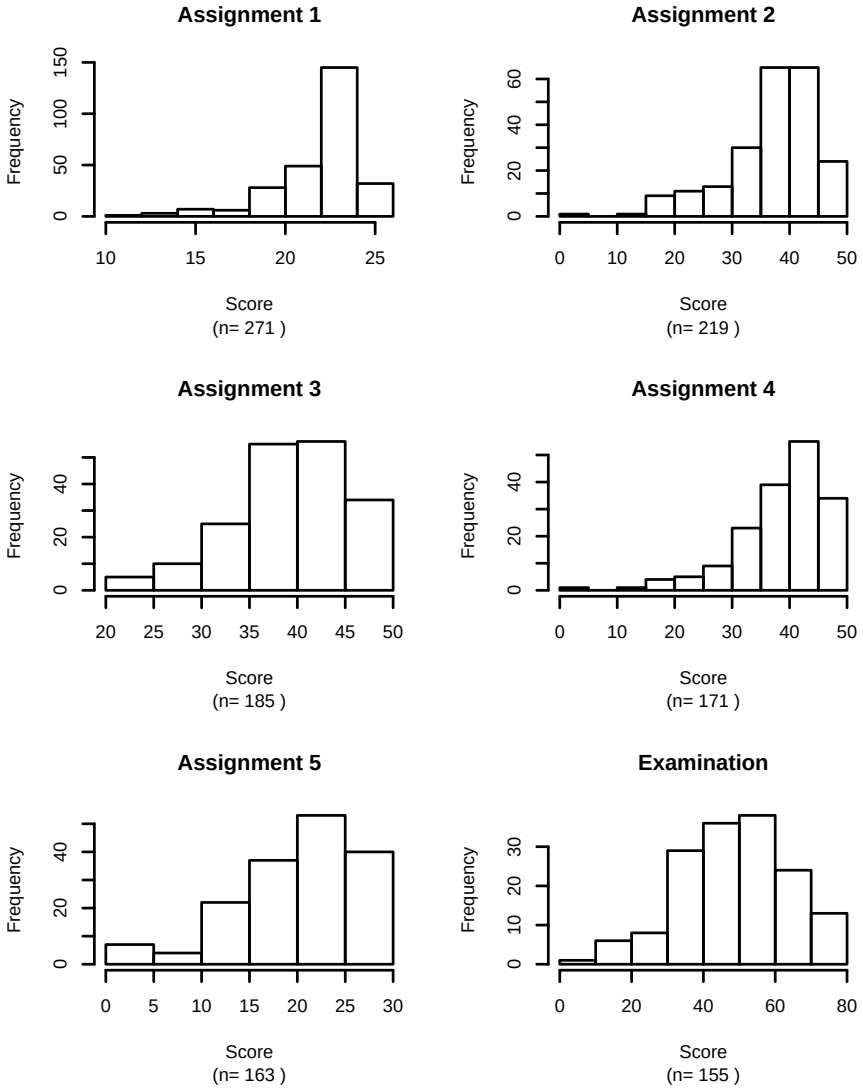

Figure 6.1: Distribution of scores for each of the assessment instruments used

p. 112]) indicated that the widespread practice of excluding such data may produce biased model estimates. Research presented in Chapter 3 suggested that the Tobit and Tweedie models could provide an alternative and novel method for modelling such performance data. Unfortunately suitable measures to predict the performance of these students were not available. This itself, is an important area for future research. There is a body of literature that deals with factors that explain the progression of students at the institution level (see, for example Tinto [96]), but very little on the progression of students at the course level. Why, for example, do some students withdraw from a course without completing any measure of performance, while others remain in the course for a longer period of time? When such research is undertaken, then methodological issues such as how best to deal with such data, can and should be explored.

References

[1] K.S. Aiken and S.G. West. *Multiple regression: testing and interpreting interaction*. SAGE Publications, Newbury Park, 1991.

[2] D. Allen. Desire to finish college: an empirical link between motivation and persistence. *Research in Higher Education*, 40(4):461–485, 1999.

[3] L. Andrews, P. Aungles, S. Baker, and A. Sarris. Characteristics and performance of higher education institutions. Technical report, Australian Department of Employment, Education, Training and Youth Affairs, 1997.

[4] P.C. Austin, M. Escobar, and J.A. Kopec. The use of the Tobit model for analyzing measures of health status. *Quality of Life Research*, 9:901–910, 2000.

[5] R. Awang-Hashim, H.F. O'Neil, and D. Hocevar. Ethnicity, effort, self-efficacy, worry, and statistics achievement in Malaysia: a construct validation of the state-trait motivation model. *Educational Assessment*, 8(4):341–364, 2002.

[6] A. Bandura. *Self-Efficacy: The Exercise of Control*. W.H.Freeman, 1997.

[7] J.P. Bean and B.S. Metzner. A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, 55(4):485–540, 1985.

[8] I. Blackman. A predictive model identifying latent variables, which influence undergraduate student nurses' achievement in mental health nursing skills. *International Education Journal*, 2(4):53–64, 2001.

[9] I. Blackman and I.G.N. Darmawan. Graduate-entry medical student variables that predict academic and clinial achievement. *International Education Journal*, 4(4):30–41, 2004.

[10] B. Bloom, M. Englehart, E. Furst, W. Hill, and D. Krathwohl. *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longmans, New York, 1956.

[11] G.W. Bohrnstedt and T.M. Carter. Robustness in regression analysis. *Sociological Methodology*, 3:118–146, 1971.

[12] W. Bosshardt. Student drops and failure in principles courses. *Journal of Economic Education*, 35(2):111–128, 2004.

[13] T. Bouffard, J. Boisvert, C. Vezeau, and C. Larouche. The impact of goal orientation on self-regulation and performance among college students. *British Journal of Psychology*, 65:317–326, 1995.

[14] G.H. Bower. Mood and memory. *American Psychologist*, 36(2):129–148, 1981.

[15] R. Breen. *Regression models: censored, sample selected, or truncated data*. Number 111 in Quantitative Applications in the Social Sciences. SAGE Publications, Thousand Oaks, 1996.

[16] M.W. Browne. Covariance structures. In D.M. Hawkins, editor, *Topics in applied multivariate analysis*, pages 72–141. Cambridge University Press, 1982.

[17] J. Burkey and T.R. Harris. Modeling a share or proportion with logit or Tobit: the effect of outcommuting on retail sales leakages. *The Review of Regional Studies*, 33(3):328–342, 2003.

[18] A. Cabrera, A. Nora, and M. Castaneda. College persistence: structural equation modelling test of an integrated model of student retention. *Journal of Higher Education*, 64(2):123–139, 1993.

[19] R. Cantwell, J. Archer, and S. Bourke. A comparison of the academic experiences and achievement of university students entering by traditional and non-traditional means. *Assessment and Evaluation in Higher Education*, 26(3):221–234, 2001.

[20] C.S. Carmichael, P.K. Dunn, and J.A. Taylor. Identifying at-risk students: Is it possible in a tertiary preparation course for adults? In P. Grootenboer, R. Zevenbergen, and M. Chinnappan, editors, *Identities, Cultures and Learning Spaces: proceedings of 29th annual conference of MERGA*, volume 1, pages 107–114. Mathematics Education Research Group of Australasia, MERGA, 2006.

[21] C.S. Carmichael and R. St. Hill. Towards quality in multiple-choice assessment. *International Journal of Business and Management Education*, 13(8):33–47, 2006.

[22] C.S. Carmichael and J.A. Taylor. The analysis of student beliefs in a tertiary preparatory mathematics course. *International Journal of Mathematical Education in Science and Technology*, 36(7), 2005.

[23] M.M. Chemers, Hu Li-tze, and B.F. Garcia. Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology*, 93(1):55–64, 2001.

[24] Chih-Ping Chou and P.M. Bentler. Estimates and tests in structural equation modelling. In *Structural equation modelling, concepts, issues and applications*. SAGE Publications, Thousand Oaks, 1995.

[25] W.S. Cleveland and S.J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the Americal Statistical Association*, 83(402):596–610, 1988.

[26] G. Considine and G. Zappala. Factors influencing the educational performance of students from disadvantaged backgrounds. In T. Eardley and B. Bradbury, editors, *Competing visions: refereed proceedings of the national social policy conference 2001*. UNSW, 2002.

[27] L. Corno. The best-laid plans: modern conceptions of volition and educational research. *Educational Researcher*, 22(2):14–22, 1992.

[28] M.S. De Berard, G.I. Spielmans, and D.C. Julka. Predictors of academic achievement and retention among college freshman: a longitudinal study. *College Student Journal*, 38(1):66–81, 2004.

[29] Alexandre de Bustamante Simas. *betareg: Beta Regression*. R package version 1.1.

[30] A. Dispeth. The relationship between intelligence, approaches to learning and academic achievement. *Scandinavian Journal of Educational Research*, 46(219–230), 2002.

[31] A.J. Dobson. *An introduction to generalised linear models*. Chapman and Hall, 1990.

[32] Peter Dunn. *tweedie: Tweedie exponential family models*, 2004. R package version 1.02.

[33] P.K. Dunn. *Likelihood-based inference for Tweedie exponential dispersion models*. PhD thesis, University of Queensland, 2001.

[34] P.K. Dunn and G.K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

[35] C.S. Dweck and E.L. Leggett. A social-cognitive approach to motivation and personality. *Psychological Review*, 95(2):256–273, 1988.

[36] B. Efron and R.J. Tibshirani. *An introduction to the Bootstrap*. Chapman and Hall, 1993.

[37] L. Elton. Are UK degree standards going up, down or sideways? *Studies in Higher Education*, 23(1):35–43, 1998.

[38] S.L.P. Ferrari and F. Cribari-Neto. Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

[39] A. Fielding, Min Yang, and H. Goldstein. Multilevel ordinal models for examination grades. *Statistical Modelling*, 3:127–153, 2003.

[40] C. Fornell and J. Cha. Partial least squares. In Richard Bagozzi, editor, *Advanced methods of marketing research*, pages 52–78. Blackwell Publishers, Oxford, 1994.

[41] K. Hall and P. Marchant. Predictors of the academic performance of teacher education students. *Research in Education*, 63:89–99, 2000.

[42] J. Hardin and J. Hilbe. *Generalized linear models and extensions*. Stata Press, Texas, 2001.

[43] J.L. Higbee and P.V. Thomas. Affective and cognitive factors related to mathematics achievement. *Journal of Developmental Education*, 23(1):8–15, 1999.

[44] Li-tze Hu, P.M Bentler, and Y. Kano. Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, 112(2):351–362, 1992.

[45] R. James, C. McInnis, and M. Devlin. Assessing learning in Australian Universities. Technical report, Centre for the Study of Higher Education, 2002.

[46] B. Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society B*, 49(2):127–162, 1987.

[47] B. Jørgensen. *The theory of dispersion models.* Chapman and Hall, 1997.

[48] M.W. Julian. The consequences of ignoring multilevel data structures in non-hierachical covariance modeling. *Structural Equation Modeling*, 8(3):325–352, 2001.

[49] Y. Kana, M. Berkane, and P. Bentler. Covariance structure analysis with heterogeneous kurtosis parameters. *Biometrika*, 77(3):575–585, 1990.

[50] K.G.Joreskog and H.Wold. The ML and PLS techniques for modelling with latent variables. In J.Tinbergen, D.W.Jorgenson, and J.Waelbroeck, editors, *Systems under indirect observation*, pages 263–270. North Holland Publishing Company, Netherlands, 1982.

[51] R. Kieschnick and B.D. McCullough. Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling*, 3:193–213, 2003.

[52] L.C. Koke and P.A. Vernon. The Sternberg Triarchic Abilities test (STAT) as a measure of academic achievement and general intelligence. *Personality and Individual Differences*, 35:1803–1807, 2003.

[53] S. Kotsiantis, C. Pierrakeas, and P. Pintelas. Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18:411–426, 2004.

[54] R. Kruse and K.D. Meyer. *Statistics with vague data*. Theory and decision library: mathematical and statistical methods. D. Reidel Publishing Company, Dordrecht, Holland, 1987.

[55] A.M. Lane, R. Hall, and J. Lane. Self-efficacy and statistics performance among sports studies students. *Teaching in Higher Education*, 9(4):435–448, 2004.

[56] J. Lane and A. Lane. Self-efficacy and academic performance. *Social Behaviour and Personality*, 29(7):687–694, 2001.

[57] Learning and Teaching Support Network. Diagnostic testing for mathematics. Technical report, LTSN, nd.

[58] J.S. Long. *Regression models for categorical and limited dependent variables*. Advanced quantitative techniques in the Social Sciences Series. SAGE Publications, Thousand Oaks, 1997.

[59] Access Economics Pty. Ltd. Review of higher education outcome performance indicators. Technical report, Australian Department of Education, Science and Training, 2005.

[60] H. MacGillivray and I. Turner. Components of learning and assessment in linear algebra. In M. Bulmer, H. MacGillivray, and C. Varsavsky, editors, *Proceedings of Kingfisher Delta'05*, pages 74–81, 2005.

[61] C. Masui and E. De Corte. Learning to reflect and to attribute constructively as basic components of self-regulated learning. *British Journal of Educational Psychology*, 75:351–372, 2005.

[62] R.R. McCrae and P.T. Costa. Towards a new generation of personality theories: Theoretical contexts for the five-factor model. In J.S. Wiggins, editor, *The five-factor model of personality: Theoretical perspectives*, pages 51–87. Guidford, 1996.

[63] P. McCullagh and J.A. Nelder. *Generalised linear models*. Chapman and Hall, London, 1989.

[64] C. McDonald and J.A. Taylor. Preliminary investigation into aggegating grades in a first year course with mixed assessment types. In M. Bulmer, H. MacGillivray, and C. Varsavsky, editors, *Proceedings of Kingfisher Delta'05*, pages 88–93, 2005.

[65] R. McDonald. Path analysis of composite variables. *Multivariate Behavioural Research*, 31(2):239–270, 1996.

[66] K. McKenzie, K. Gow, and R. Schweitzer. Exploring first-year academic achievement through structural equation modelling. *Higher Education Research and Development*, 23(1):95–112, 2004.

[67] T. Micceri. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105:156–166, 1989.

[68] H.F. Middleton and R.M. Gillies. Aptitude testing and prediction of school achievement for years 11 and 12. *Australian Journal of Career Development*, 9(2):20–26, Winter 2000.

[69] A. Miller. Personality types, learning styles and educational goals. *Educational Psychology*, 11(3):217–239, 1991.

[70] J. Multon, S.D. Brown, and R.W. Lent. Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, 38(1):31–38, 1991.

[71] H. Nevitt and G.R. Hancock. Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural equation modeling*, 3(3):353–377, 2001.

[72] N.T. Nguyen, L.C. Allen, and K. Fraccastoro. Personality predicts academic performance: exploring the role of gender. *Journal of Higher Education Policy and Management*, 27(1):105–116, March 2005.

[73] U.H. Olsson, T. Foss, S.V. Troye, and R.D. Howell. The performance of ML, GLS and WLS estimation in structural equation modeling under conditions of mis-specification and non-Normality. *Structural equation modeling*, 7(4):557–595, 2000.

[74] Y. Ommundsen. Implicit theories of ability and self-regulation strategies in physical education classes. *Educational Psychology*, 23(2):141–157, 2003.

[75] S original by Terry Therneau and imported by Thomas Lumley. *survival: Survival analysis, including penalised likelihood.* R package version 2.17.

[76] J.W. Osborne. Prediction in multiple regression. *Practical Assessment, Research and Evaluation*, 7(2):13–20, 2000.

[77] F. Pajeres and L. Graham. Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology*, 24:124–139, 1999.

[78] F. Pajeres and M.D. Miller. Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, 86(2):193–203, 1994.

[79] K.V. Petrides, T. Chamorro-Premuzic, N. Frederickson, and A. Furnham. Explaining individual differences in scholastic behaviour and achievement. *British Journal of Educational Psychology*, 75(2):239–255, 2005.

[80] M. Pokorny and H. Pokorny. Widening participation in higher education: student quantitative skills and independent learning as impediments to progression. *International Journal of Mathematical Education in Science and Technology*, 36(5):445–467, 2005.

[81] P.R. Printrich and E.V. De Groot. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1):33–40, 1990.

[82] QSA. Attrition and persistence of first-year tertiary students in Queensland: Longitudinal research study. Technical report, Queensland Studies Authority (QSA), 2004.

[83] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[84] S.B. Robbins, H. Le, D. Davis, K. Lauver, R. Langley, and A. Carlstrom. Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130(2):261–288, 2004.

[85] A. Sen and M. Srivastava. *Regression analysis: theory, methods and applications*. Springer, 1990.

[86] I. Shaw, D.P. Newton, M. Aitkin, and R. Darnell. Do OFSTED inspections of secondary schools make a difference to GCSE results? *British Educational Research Journal*, 29(1):63–75, 2003.

[87] R.M. Smith and P.A. Schumacher. Predicting success for actuarial students in undergraduate mathematics courses. *College Student Journal*, 39(1):165–177, 2005.

[88] G.K. Smyth. Using the EM algorithm to predict first year performance. *Australian Journal of Education*, 34(3):204–223, 1990.

[89] Gordon Smyth. *statmod: Statistical Modeling*, 2004. R package version 1.1.0.

[90] T.A.B. Snijders and R.J. Bosker. *Multilevel Analysis: An introduction to basic and advanced multilevel analysis*. SAGE Publications, London, 1999.

[91] R.E. Snow, L. Corno, and D. Jackson. Individual differences in affective and conative functions. In D.C. Berliner and R.C. Calfee, editors, *Handbook of educational psychology*, pages 243–310. Macmillan, New York, 1996.

[92] T. Stevens, A. Olivarez, W.Y. Lan, and M.K. Tallent-Runnels. Role of mathematics self-efficacy and motivation in mathematics performance across ethnicity. *Journal of Educational Research*, 97(4):208–221, 2004.

[93] S.S. Stodolsky. Telling math: Origins of maths aversion and anxiety. *Educational Psychologist*, 20(3):125–133, 1985.

[94] A.B. Taylor, S.G. West, and L.S. Aiken. Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorised. *Educational and Psychological Measurement*, 66(2):228–239, 2006.

[95] H. Teglasi. Assessment of temperament. *ERIC Digest*, Online, 1995.

[96] V. Tinto. *Leaving College: rethinking the causes and cures of student attrition*. University of Chicago Press, 2 edition, 1993.

[97] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[98] L.M. Wolfle. The introduction of path analysis to the social sciences, and some emergent themes: an annotated bibiliography. *Structural Equation Modelling*, 10(1):1–34, 2003.

[99] P. Zeegers. Student learning in higher education: a path analysis of academic achievement in science. *Higher Education Research and Development*, 23(1):35–56, 2004.

[100] R.D. Zettle and L.L. Houghton. The relationship between mathematics anxiety and social desirability as a function of gender. *College Student Journal*, 32(1):81–87, 1998.

Appendix A

Computer code in beta simulation

```
# Code for simulating educational performance data using the beta
distribution

# # # Created by Colin Carmichael

# Last modified:17/8/06 #
##------------------------------------------------
data<-read.table("final.txt",header=TRUE)

# Enter appropriate libraries
library(MASS)
library(betareg)
library(tweedie)
library(moments)

# # use existing data as starting point and select only complete
# students (type=3)
```

```
prop<-data$total/100 # convert data to a proportion
P.comp<-prop[data$type==3]
mtest<-data$mtotal[data$type==3]

# # fit a beta model to this data in order to estimate suitable
#initial parameters

bmodel<-betareg(P.comp~mtest)# fit model using betareg function
b0<-bmodel$coefficients[1]
b1<-bmodel$coefficients[2]
Phi<-bmodel$coefficients[3] # estimate of dispersion


# Measure the effectiveness of the linear model, commence by
#setting the seed in the randomization process.

set.seed(1)

# vary the sample size N manually
N<-100

# Have Q replications for each sample size
Q<-100

# create vectors to store the measures of effectiveness, skewness
# and rsquare

E<-numeric(Q) # stores effectiveness
```

```
SKEW<-numeric(Q)# stores skewness

RSQ<-numeric(Q)# stores rsquare

# commence the iteration process

for (p in 1:Q){
    M<-sample(mtest,N,replace=TRUE)# sample from the existing
    # mtest scores
    M<-sort(M)

    # use existing parameters to create the linear predictor
    eta<-b0+b1*M
    mu<-exp(eta)/(1+exp(eta))
    phi<-Phi # adjust the fit by varying Phi

   #set the parameters for the beta distribution
    alpha<-mu*phi
    beta<-phi*(1-mu)

    # generate a vector of response variables with suitable
    #skewness

    P<-numeric(N)
    for(i in 1:N){
        P[i]<-rbeta(1,alpha[i],beta[i])
    }

# measure the effectiveness of the linear model
```

```
    Lmodel<-lm(P~M)#create a linear model
    Yorg<-exp(b0+b1*M)/(1+exp(b0+b1*M))# then a beta model

    # calculate upper and lower confidence intervals for
    # linear model
    Ylower<-predict.lm(Lmodel,interval=c("confidence"))[,2]
    Yupper<-predict.lm(Lmodel,interval=c("confidence"))[,3]
    #calculate the effectiveness in each case.
    E[p]<-length(M[Yorg<Yupper& Ylower<Yorg])/length(M)
    SKEW[p]<-skewness(P)
    RSQ[p]<-summary.lm(Lmodel)$r.squared
}
 mean(E)
 sd(E)
 mean(SKEW)
 sd(SKEW)
 mean(RSQ)
 sd(RSQ)
```