7-12-2014

Bayesian Partially Ordered Probit and Logit Models with an Application to Course Redesign

Xueqi Wang

Xueqin Wang

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Michael Sonksen , Chairperson

Kristin Umland

Erik Erhardt

James Selig

Bayesian Partially Ordered Probit and Logit Models with an Application to Course Redesign

by

Xueqin Wang

B.A. Management, Henan Normal University, 1996
M.A. Philosophy, Nanjing University, China, 2001
M.S. Statistics, University of New Mexico, 2008

DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

May 14, 2014

Dedication

To my parents, grandpa, and grandma for raising and loving me, and giving me the values and qualities in my childhood which enable me to go through a long way; my dear children, for loving mom so much and beautiful music; and my husband, for helping with children when I took the comprehensive exam and wrote my dissertation.

Acknowledgments

I would first like to thank my advisors, Prof. Michael Sonksen, and Prof. Kristin Umland, for their kindly taking me as their student, their smart idea of the new models and this great project, and guiding and helping me in every step, for their tremendous and constant support through the difficult process of writing this dissertation. It is their personal virtue and support that strengthened me through many struggling moments and moved forward in writing the dissertation and searching for a proper job. Thank you guys for turning the hard process of my working on this dissertation into an excellent learning experience with productivity, achievement, and joy. Without my advisors, there would not be this dissertation to be presented here.

I would also like to thank Prof.Erhardt for his excellent suggestion with the proposal and the model, and making the arrangement for me to work with Prof.Umland and Prof.Sonksen. Thank Prof.James Selig for kindly suggestion and constructive comments to this dissertation. Meanwhile, I'd also like to thank Prof. Yan Lu getting me back into the program and precious encouragement and help, thank Prof. Christensen for the great classes in preparing us for the comprehensive exam, and for all of your kindness and help. Thank Dr. Winston Crandall taught my first statistics course, which opened the door for me to get my PhD in statistics. I'd like to thank everyone who worked before and who are working here now in the department, it is your hard work made my graduate school time at UNM an sweet memory.

I am also thankful for my good friends: Kim, Alvavo, Diana, Dan, Osorio, Han, Pedro, Klaus, Mohammad, Yong, Yan, Fletcher, Maozhen, Xichen, Aiqin, Yingjie, Sherry. . . , and many others I can't list.

Bayesian Partially Ordered Probit and Logit Models with an Application to Course Redesign

by

Xueqin Wang

B.A. Management, Henan Normal University, 1996

M.A. Philosophy, Nanjing University, China, 2001

M.S. Statistics, University of New Mexico, 2008

Ph.D. Statistics, University of New Mexico, 2014

Abstract

Large entry-level courses are commonplace at public 2- and 4-year institutions of higher education (IHEs) across the United States. Low pass rates in these entry-level courses, coupled with tight budgets, have put pressure on IHEs to look for ways to teach more students more effectively at a lower cost. Efforts to improve student outcomes in such courses are often called "course redesigns." The difficulty arises in trying to determine the impact of a particular course redesign; true random-controlled trials are expensive and time-consuming, and few IHEs have the resources or patience to implement them. As a result, almost all evaluations of efforts to improve student success at scale rely on observational studies. At the same time, standard multilevel models may be inadequate to extract meaningful information from the complex and messy sets of student data available to evaluators because they throw away information by treating all passing grades equally. We propose a new Bayesian approach that keeps all grading information: a partially ordered multinomial probit model

with random effects fit using a Markov Chain Monte Carlo algorithm, and a logit model that can be fit with importance sampling. Simulation studies show that the Bayesian Partially Ordered Probit/Logit Models work well, and the parameter estimation is precise in large samples. We also compared this model with standard models considering Mean Squared Error and the area under the Receiver Operating Characteristic (ROC) curve. We applied these new models to evaluate the impact of a course redesign at a large public university using the students' grade data from the Fall semester of 2012 and the Spring semester of 2013.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Intermediate Algebra is an entry-level mathematics course at a large public university in the southwestern United States, which we call the University, and enrolls approximately 2500 students per year. This course is a pre-requisite to all mathematics courses that satisfy university-level general education requirements, and successful completion of the course is considered a gateway to further success in higher education. However, Intermediate Algebra has had a historical pass rate of about 45%, and as such, has been a barrier to graduation for many students. Before the Fall semester of 2012, students enrolled in one of approximately 20 lecture-based, 60-student sections that were taught almost exclusively by part-time instructors each semester. Over time, individual instructors have tried different approaches to teaching the course in order to improve student performance, including a completely online section, online homework sets using commercial software, and integrating individual or group work into the lecture via handouts or a workbook. Because no formal evaluations of these alternative methods of instruction have taken place, no attempted changes have been widely implemented. In the Spring of 2012, a team of University administrators, faculty,

and graduate students researched different instructional approaches used at other institutions with the intent of increasing the success rate of students in both Intermediate Algebra and the subsequent courses that students need to take to graduate from the University. The team determined that there was significant promise in one particular model and proposed to pilot it in the Fall of 2012. Enthusiasm for the redesign amongst university administrators accelerated the full implementation, which occurred in Spring 2013. The Redesigned course is a self-paced course offered in a computer lab. The model employs mastery-learning, which allows students to move quickly through material they already understand and to spend more time working on material that is harder for them, receiving help from circulating tutors when needed. In the Fall of 2012, 1308 students enrolled in Intermediate Algebra. Of those, 1092 enrolled in a traditional 60-student lecture, and 216 enrolled in the Redesigned course.

The way that students enroll in university courses, coupled with the complexity of getting permission to do experiments on students, means that it was not possible to randomly assign students into the two different instructional groups. Thus, this evaluation is an observational study that models student success using student and instructor covariates and using several different model structures and outcome measures. The two instructional methods of interest will be called "Traditional lecture" and "Redesign" for the purpose of this dissertation.

The primary research questions for the evaluation component are:

1. What is the impact of the Intermediate Algebra redesign on
   (a) Student success in Intermediate Algebra in the semester taken,
   (b) Student success in mathematics courses taken in subsequent semesters that satisfy general education requirements at the University?

Additionally, the primary research questions related to the statistical modeling are:

2. How will advanced (novel) statistical modeling improve our understanding of the dif-

ferential impact of the Traditional lecture and Redesign instructional methods?

3. What are good model structures for analyzing this kind of data?

1.2 Project Summary

In 1994, Robert C. Heterick, Jr., former president of Educom, pointed out that our knowledge about how using information technology in teaching and learning affects learning outcomes and cost was still meager, and one of our continuing tasks must be to measure, hypothesize, and finally formalize theories about how technology applies to education. One of our great failings so far as a community was relying too heavily on the anecdotal and not doing the hard work to justify our concepts through meticulous measurement and theory building. He also addressed the need to help faculty and learners measure the outcome of incorporating information technology to teaching (Heterick, 1994). While he made these statements 20 years ago, they are still important today.

In response to the great need for college-level math course evaluation using advanced statistical tools, we developed two new statistical models: Bayesian partially ordered probit and logit models to help measure student outcomes more accurately despite the limited information we have, because usually students' success is recorded as letter grades.

We started with standard statistical approaches like obtaining descriptive statistics of the data and classical ANOVA (analysis of variance), t-test, and pass rate calculations that have been used by other authors. We also used standard models like multilevel logit and probit models to check if there is a significant difference in student mean passing rates between the Traditional lecture and the redesigned Intermediate Algebra course. Due to the limitation of the standard models to our data, we developed and implemented new statistical models: Bayesian partially ordered probit/logit models. The Bayesian partially ordered probit/logit models are the statistical contribution of this dissertation, and they are shown to be superior

to standard models such as a multilevel probit model and an ordered probit model with a five-fold cross validation based on Mean squared Error and Area Under the ROC curve. This will be discussed in Chapter 3. Our models give a more precise prediction than the standard models do for partially ordered letter grade data, which is one of the most frequently encountered data forms used to measure educational outcomes.

Our models take into account independent variables such as gender, high school graduate or GED certificate holder, the amount of time after high school graduation, ACT and SAT math scores, semester credit hour load, semester GPA (a weighted average of students' grade in the semester while taking Intermediate Algebra), race/ethnicity, and include a random effect. We use instructor as a second-level variable to model the data in Fall 2012, because students are nested under instructors, and one instructors taught only the Redesigned sections, while all other instructors taught the Traditional lecture sections. The models are based on the assumption that student letter-grades are based on an unobserved score between 0 and 100, and they predict students' course scores out of 100 that we do not observe. We provide these novel statistical tools to help more accurately evaluate the course redesign.

This research started out as an evaluation of a particular implementation of a technology-based course redesign, but it has implications that reach beyond the particular implementation studied. Even more than the findings, the methods employed in this research can be applied to improve the evaluation of similar efforts elsewhere. In fact, these models can be used to analyze any partially ordered data in fields besides education. We applied these models to analyze the performance of Intermediate Algebra course redesign at the University. After finishing this dissertation, we will work on incorporating these new statistical models into SPSS, a spreadsheet-based software which anyone with knowledge of Excel can use. The advantage of coding the models into SPSS is that someone with only a little bit of statistical knowledge can apply the models to their partially ordered data. This will make it possible for educators to easily use it to analyze their data more precisely than with standard models. These models fill the void of measurement work: accurately evaluating course

redesign despite the limited information we have.

Chapter 2

Course Redesign and Assessment: Background and Implementation Structure

This chapter contains four sections. The first provides background information on teaching Intermediate Algebra. The second section defines the terms we use in this dissertation about course redesign. The third section introduces the nature of the Intermediate Algebra course redesign at the University. The fourth section reviews the methods currently described in the education literature for the evaluation of course redesign.

2.1 Background

The poor quality of mathematics education in the United States at the level of Intermediate Algebra was known as early as the late 1970s and early 1980s. Gardner et al. (1983) found that high school curricula no longer had a central purpose that would give students systematic training. Only 31 percent of recent high school graduates completed Algebra II,

the rough high school equivalent of Intermediate Algebra. They also found that time spent on homework and average student achievement had declined, despite the fact that grades were rising. As a comparison, in many other industrialized countries, courses in mathematics (other than arithmetic and general mathematics) started in grade 6 and are required of all students, and students spent about three times the number of class hours in math as even the most science-oriented U.S. students. At that time, there were thirty-five states which required only 1 year of mathematics in high school (Gardner et al., 1983).

This situation has remained largely unchanged. In the report from National Mathematics Advisory Panel in 2008 (Panel, 2008), it was found that 15-year-old students in the United States ranked 25th among their peers in 30 developed nations in math literacy and problem-solving. Failure to thrive in mathematics during their K-12 education causes difficulty and stress once students go to college. According to the report of National Mathematics Advisory Panel in 2008, which was adopted by Department of Education Secretary Margaret Spellings, "The sharp falloff in mathematics achievement in the U.S. begins as students reach late middle school, where, for more and more students, algebra course work begins." Moreover, they found that students who completed Algebra II in high school were more than twice as likely to graduate from college compared to students with less mathematical preparation (Panel, 2008). Because the content of Algebra II in high school mentioned in the report is roughly equivalent to the content of Intermediate Algebra in college, we can see that students who are placed into Intermediate Algebra in college (because they did not complete or did not master Algebra II in high school) are at a greater risk of failing out of college. Intermediate Algebra prepares students to succeed in College Algebra, the gateway course to most Math, Science, and Engineering degrees. So the quality of the preparation that students receive in Intermediate Algebra impacts their ability to graduate with a degree in a technical field. Small (2002) reported that the failing (FDW) rate in College Algebra is often in the 40%-60% range and "College Algebra blocks academic opportunities and plans for approximately 200,000 students per semester." Moreover, failing in College Algebra or Intermediate Algebra is a common cause of students dropping out of college. At many institutions, nearly half of

these non-passing students drop out of college entirely (Beaudrie, 2002). In order to build up students' math skills and help students pass College Algebra, universities offer a lower level math course: Intermediate Algebra. The level of the mathematics course that students are placed in the first year of their college is typically decided by their ACT, SAT or a math placement test score.

When students drop out of college, there is a negative long-term effect on both the students and on the country's economy. To students, higher education can bring them a well-paid job and a better future, and dropping out of college takes away many possible opportunities from them. In addition, more highly educated workers are needed for the economy in many fields. The workforce may be in short supply of highly educated people when many students cannot finish their college. Improving learning and achievement in Intermediate and College Algebra can have a positive impact on both students and the economy. The unacceptable failing rates, the significant economic impact, and student frustration have motivated mathematics educators to implement novel approaches intended to improve students performance in remedial and entry-level college math courses such as Intermediate Algebra and College Algebra.

Technological developments in the last three decades have brought new options for reforming the traditional teaching model. Computer software programs for courses in many subjects have been developed and are in use at many institutions. For example, early in Fall 2000, the University of North Florida invested in a 40-seat College Algebra lab and added a no-credit lab hour in addition to the traditional three hours per week of class time. The result was that the passing rate of students who used the computer lab increased 12% compared to the students who did not use the computer lab (Dedeo, 2001). More advanced computer-based college-level mathematics courses have been implemented since then in many universities and received encouraging results. At Mississippi Valley State University, students meet 75 minutes per week in class with an instructor and have 75 minutes per week scheduled time in a lab, with a total of 3 hours per week in lab required, and take exams online (except

the final exam). Mississippi Valley State University offered fewer sections after the redesign because there were fewer repeats. Jackson State Community College required that a student scheduled 3 hours per week in a lab, instead of a regular class, exams were offered online and proctored; the pass rate was 42% and improved to 59% for Basic Math. At the University of Central Florida, students had one class per week with an instructor, an additional 3 hours of lab time, online exams, and their 65% pass rate went up to 74% (Mason et al., 2012).

2.2 Definition of Terms

Before we describe the introductory-level college math course redesign studied here, there are some terms we need to clarify.

2.2.1 Emporium Model

The Emporium Model is a teaching model in which students use computer-based resources to learn and be assessed. It is also called the Computer Assisted Instruction (CAI) model, and has generally been accepted by educators and students as an alternative to the traditional lecture teaching model in college level introductory courses. In many universities and colleges, the emporium model has been shown to enhance students learning and reduce costs.

2.2.2 Traditional Instruction and Course Redesign

There are many possible ways to change a course structure to improve student learning. In this study, the course redesign uses modern information technology to reorganize the teaching of a large enrollment course with the intention of improving student learning outcomes at a lower cost. The course redesign is not just an online course, but uses information technology to provide more individualized instruction (Twigg, 2011). In particular, the course redesign

at the University consists of an implementation of the emporium model, which is a redesign structure that is supported by the National Center for Academic Transformation (NCAT). There are also other course redesign models such as supplemental model, replacement model, fully online model, buffet model, and the linked workshop model, and the evaluation methods presented here could be used to analyze any of these models.

The traditional approach of teaching Intermediate Algebra referred to in this study is a lecture-based approach. Under the current traditional teaching model, the Department of Mathematics and Statistics at the University offered approximately 45 lecture-based sections per year with at least 55 students each which are taught almost exclusively by part-time instructors.

2.3 The Redesigned Course Structure

The Redesigned course at the University is an implementation of the emporium model for teaching Intermediate Algebra. During Fall 2012, the University ran a pilot with 216 students, divided into two sections, which were taught by one instructor. To accomplish this, there was a computer lab with about 25 computers that was open 60 hours per week. Each student was required to spend at least 3 hours per week working in the lab. The full implementation in Spring 2013 was situated in a computer lab with 125 computers that can accommodate 1000 students for about 5.5 computer hours per week. There was also a separate room with 15 computers for taking exams. According to Mason et al. (2012) and direct communication with the course coordinator, the redesigned Intermediate Algebra course has the following characteristics:

1. **Lectures replaced with individualized study**. Lectures are replaced with required time in the computer lab. The content of the course is presented using a commercially available computer and internet learning environment called *ALEKS* (ALE, 2013).

Students work on their course content in the lab, where on-demand, personalized assistance is available through the software and lab staff.

2. **Self-paced with modularized content**. In a traditional lecture, all students work at the same pace; the redesigned Intermediate Algebra course allows students to move quickly through and test out of the material they already know so that they can spend more time on concepts and problems that are more difficult for them. At the beginning of the semester, students take an online assessment, then they are given a specific plan for individual study designed to fill in their knowledge gaps. The Intermediate Algebra content is organized into three modules. Once a student has passed a module exam at 75% proficiency, they will not be required to return to that module even if they do not finish all three modules in the first term of enrollment. Students may finish the course at any time during the semester, and may also complete the content over the next two semesters (including summer). In the full implementation starting from Spring 2013, students are also allowed to work on weekends and during breaks. The grades students receive for the redesigned Intermediate Algebra course are A, B(with +/-), C+, C , W (Withdrawal), I (Incomplete), CR (passing with credit), NC (failing with no credit) or AUD (Audit).

3. **Individualized and outcome-based instruction**. At the beginning of each module, students take a diagnostic exam that determines the individual students' strengths and weaknesses and creates an individual learning plan. Students get online help via videos, tutorials, examples, and animations on the topics which are difficult for them. There are also instructors working in the lab to help students with difficult parts of the curriculum when needed. Instructors also monitor students' progress online, and communicate with them via email to keep them on task and on time. Tutors are always available in the lab to answer questions and give individual instruction. Each students learning plan is modified based on his or her own strengths and weaknesses shown by their homework or quiz solutions. A math coach is also provided to teach student

academic and behavior skills like time management, planning, persistence, and give timely encouragement and insight.

4. **Cost saving**. At the time of the proposal, the Intermediate Algebra redesign was projected to reduce the cost per student from $82 to $64, with a potential cost-savings of about $45,000 per year (Mason et al., 2012). This is similar to many other institutions: Virginia Tech produced savings of about $53 per student (the cost went from $77 to $24). In 2000, the University of Alabama redesigned Intermediate Algebra and the redesign reduced the cost per student from approximately $122 to $86, a 30-percent savings. The University of Idaho redesigned Intermediate Algebra and Pre-Calculus in 2000, and the new active-learning model reduced the total cost of offering both courses from approximately $338,000 to $235,000, a reduction of 31 percent (Twigg, 2011).

In the Fall semester of 2012, the University ran two parallel methods to teach Intermediate Algebra: 1092 students were taught using the traditional lecture method, and 216 enrolled in one of two pilot sections of the Redesigned course. The traditional sections used a common final exam that was graded by all instructors together, with each instructor grading one or two questions for impartiality and consistency. The final exam for the redesigned sections had similar questions from the same topics as the traditional lecture sections, but students in the redesigned sections took the final exam online with questions generated by the computer software from the same topics. Students in the redesigned sections are permitted to repeatedly take the final exam, but the questions are different each time they try. Likewise, the final exam for each student is different but comprised of similar types of questions from the same topics. This ensures that students who finished the course before the end of the semester could not release the questions to the ones who finished later. In Spring 2013, all students were enrolled in the Redesigned course.

2.4 Statistical Evaluations of Course Redesigns

Heterick (1994) brought up the importance of accurately measuring the impact of course redesigns. In the past, the statistical evaluation of course redesigns was limited to pass-rate calculations, standard t-tests, ANOVA or ANCOVA models. Besides these basic calculation and analysis with standard models, a complete evaluation of the performance of the Intermediate Algebra redesign using advanced statistical tools has the potential to provide a more precise picture of the impact of a course redesign.

Research into the impact of college-level mathematics course redesigns shows some potential promise but also leaves some unanswered questions. In this section, we will review the evidence for and against course redesigns at various institutions, and provide a critique of some of the methods of evaluation that were used. This leads us to suggest alternative models for evaluating course redesigns. In her paper "The math emporium: A silver bullet for higher education", Carol Twigg, the president and CEO of National Center for Academic Transformation (NCAT), presented the evidence of success of the math emporium model at the University of Alabama, University of Idaho, Jackson State Community College, Louisiana State University, Alcorn State University, the University of Missouri-Saint Louis, Cleveland State University, and the University of Central Florida (Twigg, 2011). After redesigning Intermediate Algebra in 2000, student success rates (grades of C or better) increased from 40.6 percent in Fall 1999 to 78.8 percent in Fall 2003 at the University of Alabama. In 2008, Alcorn State University, students in the redesigned College Algebra course using the emporium model performed significantly better than those in the traditional format: the average score on mid-term and final exams of Fall 2008 traditional sections was 55.89, while that of Fall 2009 redesigned sections was 66.16. Student success rates in Intermediate Algebra at the University of Idaho in 2000 increased from 59 percent in the traditional format to 75 percent in the redesigned format, and in Precalculus from 68 to 75 percent. Similar exciting results for introductory-level mathematics course redesigns in the other four institutions listed above were also reported.

Hagerty et al. (2010) evaluated the effect of Algebra course redesign at Black Hills State University ($BHSU$), a four-year, public liberal arts school located in Spearfish, South Dakota from the years of 2001 to 2006. After students enrolled in eight sections of College Algebra, four sections were selected to use the same online software (the redesigned sections), and four sections were taught using the traditional approach. Instructors were assigned to a redesigned section and a traditional section for both the comparison of the effect by an individual instructor as well as the overall effect. The result showed "a 21% increase in passing rate (from 54% to 75%), a 300% increase in enrollment in the next mathematics course in the program (trigonometry), a 25% improvement in attendance, and a statistically significant increase in Collegiate Assessment of Academic Proficiency ($CAAP$, a nationally normed test) scores". The passing rate was defined by a "C" or above. What we have seen so far that the success of course redesigns is mainly from directly recorded pass rates, final exam scores and other test scores. However, the College Algebra redesign at a community college in the southeast did not go well, as reported by Wynegar and Fenster (2009). Wynegar and Fenster (2009) conducted an analysis of the impact of Computer Aided Instruction (CAI) teaching model and traditional lecture on student learning of College Algebra, where student learning was defined by grades in the College Algebra course, controlling for instructor specific grading differences. They translated the students' final exam scores on a scale where the grade "A"=4.0, "B"=3.0, "C"=2.0, "D"=1.0, and "F"=0.0 and conducted a simple one way ANOVA t-test. They found that, "students in the traditional lecture averaged a 2.07, compared to 1.61 for CAI. Students taking the course with traditional lecture outperformed, on average, those in CAI by 0.46, a difference of almost one-half a letter grade." The 0.46 difference was statistically significant with from a t-test with t=1.99, and p-value $<.05$. Considering the failing rate, "29% of students failed the CAI course, compared to 20% that failed the traditional lecture course."

This analysis has several serious defects in the methods they used:

1. The response or dependent variable must be continuous, and have normal or roughly

normally distributed residuals, which is the key assumption of the ANOVA model and t-test. In this case even though the categorical letter grades were transformed to numerical values, the response variable is still by no means continuous, nor are the residuals normally distributed, so the assumption of the t-test is violated.

2. When there are many ordered categories, a regression or ANOVA model can be fit with the transformed numerical dependent variable (Gelman and Hill, 2007), if the normal distribution assumption holds. But clearly, it is not the case with the data of Wynegar and Fenster (2009). In this case not only the assumption of ANOVA is violated, but also the continuous model does not take into account the ceiling and floor effects of the ordinal outcome (Winship and Mare, 1984). The ceiling and floor effects underestimate the effect of the treatment by putting an artificially low ceiling on the high response, and a high floor on the low response, thus shortening the distance of high responses and low responses. The transformed variable means that A (4.0) is twice as good as C (2.0), four times good as D(1.0), this is not a fact. How were other grades such as CR (pass with credit), NC (fail with no credit), W (Withdrawal), I (Incomplete), AUD (Audit) treated? The authors did not mention this in the paper. These grades cannot be properly represented with one number because we do not have much information about them, and it is also inappropriate to remove them from the study as they encode important information about student success.

3. Last, since the paper never mentions that students were randomly assigned to the different treatment groups, this is an observational study. The effect of confounding variables on the response (students' College Algebra grade) such as students' math preparation level, family social economic status, and high school (or high school GPA) should be considered. The impact of these variables on the students' College Algebra performance between the groups might not be equal since the students were not randomly assigned. The authors did not consider the different influences of other covariates on the response between the traditional lecture and CAI.

Due to these reasons, the analysis of Wynegar and Fenster (2009) does tell us some information about their study, but their conclusion for this study is suspect. We have seen similar practices in education literature: researchers simply match letter grades with a numerical value (the GPA that each letter grade indicates), and fit a regression or ANOVA model. Stephens and Konvalina (1999) investigated whether the integration of the computer software MAPLE into Intermediate and College Algebra would have an effect on students' performance in the courses. They used one-way ANOVA and boxplots to compare the exam scores of the experimental group (with computer software used in teaching the course) and the control group (without using computer software). A good thing about their study is that students from both groups took the same final exam, and the exam score is known. Their conclusions were that the mean scores for both Intermediate Algebra and College Algebra in the experimental group is higher than those of the control group, but the test is not significant because the p-value is big. Stephens and Konvalina (1999) did not consider the randomization principle for designing an experiment, and did not include the influence of the difference in students' information, such as students' background information, math preparation level before taking the Intermediate Algebra and College Algebra between the two groups either. Hence, the conclusions are not very convincing. This is a common mistake in evaluating a course redesign: assuming that the students are randomly assigned into the two groups, traditional lecture sections and redesigned sections, and assuming the students' backgrounds and life situations in the two groups are the same. In most cases, evaluations are observational studies because it is expensive and time consuming, so often is not possible, to randomly assign students into the control and treatment groups. Students typically take several classes, and many have work or family obligations, so we allow each student to register in the class that fits his or her schedule. This means that students' conditions that affect their success in the course might be different between the traditional sections and the redesigned sections; thus, this "background" difference must be accounted for when we evaluate student success between the two groups.

The next two analyses did not adequately consider the difference in student success

caused by confounding covariates either. In Spring 2010, Bishop (2010) analyzed the effect of computer-based instruction on student mathematics achievement and students' attitude toward mathematics in developmental and introductory mathematics courses: Elementary Algebra, Intermediate Algebra, and College Algebra, at a community college. Her participants were 112 students enrolled in 6 algebra classes. The control group was the three classes taught using traditional lecture instruction, and the treatment group was the three classes taught using computer-based instruction via the interactive online software $MathXL$. The control and treatment groups were taught the same objectives and received instruction two days a week for 75 minutes per day. The ANOVA results were that students in the traditional lecture group had significantly higher score than students in the computer-based classes. ANCOVA results of the pre- and post-tests showed no significant difference in the achievement between the control and treatment groups.

The "modified emporium model" experiment for teaching College Algebra went better at the University of Missouri-Kansas City(UMKC) (Brown, 2012). The study at UMKC used all 193 students enrolled in College Algebra in University of Missouri-Kansas City in 2011-2012 academic year, with 87 students in the redesigned College Algebra course and 106 students in the traditional lecture. Brown (2012) analyzed the difference in mean final exam score and failing (DWF) rate between the students enrolled in the redesigned and traditional lecture sections, and also compared the difference of mean final exam score between the two teaching models across different gender and race/ethnicity, as well as within each gender and each race/ethnicity. His analysis was conducted by Statistical Package for the Social Sciences (SPSS) using t-tests, two-proportion z-tests, Multivariate Analysis of Variance (MANOVA) and ANOVA. The result was that performance for the redesign students was slightly higher (59.14 versus 57.93), but the difference was not significant at an overall level of significance of 0.05 (the p-value = 0.358); although they hoped that the DFW rate would be lower for the redesign group than for the traditional group, the fact was that a 41% DFW rate in the redesign and a 21% DFW rate in the traditional approach. Other results are: males scored higher than females in the redesign (59.6 versus 58.4) while

females scored higher than males with the traditional approach (59.6 versus 55.0). The final exam scores between redesign and traditional lecture students are: African-American students (43.6 versus 43.3), Caucasian students (65.8 versus 62.1), Hispanic students (76.25 versus 45.9), and non-residential, international students (84.8 versus 77.8). However, this difference in students' College Algebra score attributed by race/ethnicity of students was not included in their models. Brown (2012) concluded that the slightly better performance on the final exam in the Redesign sections at UMKC could be considered quite promising for future implementation of the modified emporium model, especially considering that the implementation in the Spring semester of 2012 was the first implementation and the students in the Redesign sections were less prepared than those in the traditional lecture sections as measured by student ACT, SAT and high school GPA. If Brown (2012) had incorporated the demographic information of students' as well as their math preparation level into a regression model, to include the influence of this information on students' course score, the evaluation would have been more accurate. Bishop (2010) and Brown (2012) had good response data: students' exam score at the end of the semester, but the information was not well utilized in their analysis.

As we can see from these examples, the statistical evaluation of college-level course redesign is often no more sophisticated than a pass-rate calculation, standard t-test, ANOVA or ANCOVA. Furthermore, many more advanced standard modeling approaches, such as multilevel probit/logit and ordered probit/logit models, have limitations as described in Chapter 4. An evaluation of the impact of the Intermediate Algebra redesign is in need of advanced statistical tools that can give a more accurate comparison than the earlier methods.

Chapter 3

A Review of Relevant Statistical Models

In this chapter, we focus on giving a summary of the development of the models for discrete and ordinal data. The models are ordinary probit/logit models, multilevel probit/logit models, and ordered probit/logit models. We review both the Bayesian and non-Bayesian version of these models, as well as their associated computation methods. A note on notation: the bold symbols represent matrices or vectors.

3.1 Models

In this section, we review Bayesian models with dichotomous or polychotomous response variables: probit/logit models, multilevel probit/logit models, and ordered probit/logit models. Non-Bayesian counterpart of these models are also discussed.

3.1.1 Probit and Logit Regression Models

A Introduction of Probit and Logit Regression Models

For illustration purposes, we start with a review of linear regression. The standard linear regression model is:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i \tag{3.1}$$

with dependent variable $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and independent variables $\boldsymbol{X}_i' = (1, X_{i1}, \ldots, X_{ip})$, n is the number of observations, $i = 1, 2, \ldots, n$ and p is the number of predictor variables. ϵ_i's are called disturbance terms or error terms.

A linear regression model assumes that the dependent variable, $\boldsymbol{Y}$, is continuous, although this does not put any restriction on the possible values that the independent variables may take on. Other assumptions with the linear regression model in Equation 3.1 are: ϵ_i is not correlated with any of the independent variables, ϵ_i has a mean of 0, $E(\epsilon_i) = 0$, and uncorrelated from one another, $Cov(\epsilon_i, \epsilon_j) = 0$ (called "serial independence"), and $Var(\epsilon_i) = \sigma_\epsilon^2$, where σ_ϵ^2 is constant across all observations (called "homoscedasticity"). These assumptions about ϵ_i are often called Gauss-Markov assumptions. Another common assumption is that ϵ_i is normally distributed. This implies that $\hat{\boldsymbol{\beta}}$ is normally distributed, so hypothesis tests and confidence intervals about $\hat{\boldsymbol{\beta}}$ can be easily constructed.

Logistic regression was established by Cox (1971) to analyze binary or dichotomous data. When the response ($\boldsymbol{Y}$) is binary or dichotomous, i.e, "Pass" or "Fail", "Yes" or "No", the probability of "Pass" or "Yes" can be modeled when the outcomes of $\boldsymbol{Y}$ are assumed to be mutually exclusive and exhaustive, we can use a logit or probit model to model the response data on a set of explanatory variables ($\boldsymbol{X} = (X_1', X_2', \ldots, X_n')$). A logit model is described as:

$$p_i = P(Y_i = 1|X_i), \tag{3.2}$$

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = \boldsymbol{X}_i'\boldsymbol{\beta}. \tag{3.3}$$

The probit model is an alternative inverse standard normal function with $\Phi^{-1}(\cdot)$ replaces the logit:

$$\Phi^{(-1)}(p_i) = \boldsymbol{X}_i'\boldsymbol{\beta}. \tag{3.4}$$

Where Φ is the cumulative distribution function (CDF) of the standard normal distribution, and $\Phi^{(-1)}$ is the inverse of the CDF. The functions *logit* and $\Phi^{(-1)}$ are examples of link function: functions which map the mean of the the actual binary response to the real number line. This is done so that, at some level, the model is a linear model. The choice of a link is up to the user, see McCulloch (2006) and Thompson and Baker (1981) for a discussion of link functions. We consider only these two link functions in this dissertation. Figure 3.1 displays the predicted probability by probit and logit link functions. We see that as long as the predicted probability is not extreme, and close to 0 or 1, the probability estimated with the same value for $\boldsymbol{X}_i\boldsymbol{\beta}$ but using the two different link functions are almost the same.

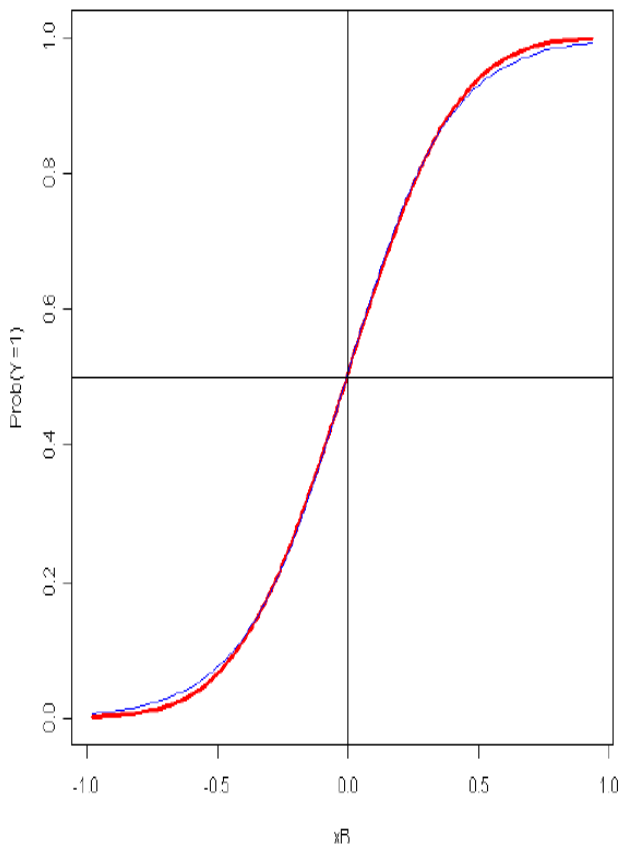

Figure 3.1: A comparison of predicted probabilities from probit and logit models. The blue curve represents the predicted probabilities from the logit model, and the red curve represents the predicted probabilities from the probit model. The vertical axis represents the probability of Y = 1, the horizontal axis represents the values taken by $\boldsymbol{X}_i'\boldsymbol{\beta}$.

The logit model is given in Equation 3.3, since the transformed response $log(\frac{p_i}{1-p_i}) = \boldsymbol{X}_i'\boldsymbol{\beta}$, if the ith element of $\boldsymbol{X}$ changes by one unit, then the logarithm of the odds ratio will change $\boldsymbol{\beta}_i$ times, because the right side will become $\boldsymbol{X}_i'\boldsymbol{\beta}_i + \boldsymbol{\beta}_i$, then taking the exponential to get rid of the logarithm, so the odds ratio will change by a factor of $e^{\boldsymbol{\beta}_i}$. Since for small values of $\boldsymbol{\beta}_i$ (for example, $\boldsymbol{\beta}_i < 1$), $e^{\boldsymbol{\beta}_i} \approx 1 + \boldsymbol{\beta}_i$, this is almost the same as saying a $\boldsymbol{\beta}_i$% increase in the odd ratio. For example, if $\boldsymbol{\beta}_i = 0.25$, then $e^{.25} = 1.28$. Thus, when X increases by one unit, the odd ratio will increase by a factor of 1.28, or increases by 28%. While the expression of probit model in Equation 3.4 depends on not just $\boldsymbol{\beta}$, but on the value of $\boldsymbol{X}_i$ and all other variables in the equation. So to even calculate the impact of $\boldsymbol{X}_i$ on Y we have to choose values for all other variables $\boldsymbol{X}_j$ variables. Typical options are to set all variables to their means or their medians. Another approach is to fix the $\boldsymbol{X}_{ij}$ and let $\boldsymbol{X}_{iJ\neq j}$ vary from its minimum to maximum values, then you can plot how the marginal effect of $\boldsymbol{X}_{ij}$ changes across its observed range of values. Thus, the simplicity of the interpretation of the logit models has lead to their popularity. For this reason, logit models for binary or polychotomous response data are preferred, for example, see O'Brien and Dunson (2004).

Estimation of $\boldsymbol{\beta}$ for Probit/Logit Regression Models

In this section, we will review the common estimation strategies for probit/logit models.

Least Square Estimation Using the explanation of Aldrich and Nelson (1984), the method of Least Square Estimation is to find parameter estimators that make the predicted $\boldsymbol{Y}$ values the closest to the actually observed $\boldsymbol{Y}$, based on the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$. In ordinary regression, this closeness is measured by the sum of squared differences between the predicted and observed Y. The Least Square Estimator for $\boldsymbol{\beta}$ is the $\hat{\boldsymbol{\beta}}$ that make the sum of squared error the smallest.

The Ordinary Least Square (OLS) estimates of $\boldsymbol{\beta}$ in linear models are those which minimize the sum of squared error $\sum_{i=1}^{N} \hat{\epsilon}^2 = \sum_{i=1}^{N} (Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}})^2$. According to Gauss-Markov

theorem, the OLS estimators are Best Linear Unbiased estimators (BLUE) (Christensen, 2011). The OLS estimates for probit or logit models becomes

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} (Y_i - \frac{exp(\boldsymbol{X}_i'\boldsymbol{\beta})}{1 + exp(\boldsymbol{X}_i'\boldsymbol{\beta})})^2 \right]$$

Horrace and Oaxaca (2006) found that OLS estimates are frequently biased and almost always inconsistent in the linear probability models. Therefore, OLS estimates of a dichotomous dependent variable are not desirable. Besides the assumptions we just discussed in Section 3.1.1 about probit or logit models, we also assume that Y_is are statistically independent of each other conditional on $\boldsymbol{X}\boldsymbol{\beta}$. Similar to OLS regression, there should not be exact linear dependency among the independent variables ($\boldsymbol{X}$) and the number of observations should be greater than the number of $\boldsymbol{X}$ variables ($n > p$), each $\boldsymbol{X}$ must have some variations, and no two or more $\boldsymbol{X}$ are perfectly correlated. Hence, probit and logit models can have the same multicollinearity problem as OLS regression does, which cause computational imprecision, unstable estimates, and large sampling error. Using weights in Ordinary Least Square estimation for probit and logit models make the statistical properties (unbiasedness, efficiency, normality) hold only asymptotically.

Maximum Likelihood Estimation In contrast to ordinary regression estimated by the method of Least Square Estimation, the standard approaches to fit a probit or logit model is to use Maximum Likelihood Estimation. The Maximum Likelihood method has a long history going back to Fisher (1922). Maximum Likelihood Estimates (MLE) are the parameter values such that the likelihood ($L(\boldsymbol{\theta}) = f(\boldsymbol{Y}|\boldsymbol{\theta})$) obtains its maximum value. MLEs enjoys a large number of properties such as consistency, asymptotic normality, and asymptotic efficiency. See Cramer (1946), and Wald (1949) for examples of such properties.

In contrast to OLS, the method of MLE has a different objective that can be described as follows. Let $p_i = P(\boldsymbol{Y}_i = 1|\boldsymbol{X}_i)$, so $P(Y_i = 0|\boldsymbol{X}_i) = 1 - p_i$, then the probability of observing outcome Y_i is $P(Y_i|\boldsymbol{X}_i) = {p_i}^{Y_i}(1 - p_i)^{Y_i}$. The probability of observing a particular sample of

N values of Y, given all N sets of of values of $\boldsymbol{X}_i$ is given by the product of N probabilities,

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{i=1}^{N} {p_i}^{Y_i}(1-p_i)^{Y_i},$$

because the observations are independent. Note that p_i is formulated in Equation 3.3 and 3.4. The value of p_i depends on $\boldsymbol{\beta}$ in the models, thus the likelihood is $P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta})$. The likelihood function $L(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X}) \equiv P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta})$. With this function, given the data on $\boldsymbol{Y}$ and $\boldsymbol{X}$, and an estimate for $\boldsymbol{\beta}$, calling it $\hat{\boldsymbol{\beta}}$, the probability or likelihood between 0 and 1 would be calculated. The principle of Maximum Likelihood Estimation is to use the parameter value of $\hat{\boldsymbol{\beta}}$ that can make the likelihood given the particular observed data as large as possible. That is, the $\hat{\boldsymbol{\beta}}$ we would take, $L(\hat{\boldsymbol{\beta}}|\boldsymbol{Y},\boldsymbol{X}) = max_{\boldsymbol{\beta}}L(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X})$, or $\hat{\boldsymbol{\beta}} = argmaxL(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X})$. In brief, the difference between Least Square Estimation and Maximum Likelihood Estimation is that Least Square Estimation picks the the parameter estimates that produce the smallest sum of squared errors in the fit between the model and data, while Maximum Likelihood Estimation chooses the parameter estimates that imply the highest probability or likelihood of having obtained the observed sample $\boldsymbol{Y}$. The method of Maximum Likelihood is widely used and can be found in most computer packages. Given a random sample of n observations, for the probit model: $P_i = P(Y_i = 1) = \Phi(\mathbf{X_i}^{\top}\boldsymbol{\beta})$, the MLE of $\boldsymbol{\beta}$ is obtained by maximizing the likelihood function:

$$L(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X}) = \prod_{i=1}^{N} \left[\Phi(\boldsymbol{X}_i'\boldsymbol{\beta})\right]^{Y_i} \left[1-\Phi(\boldsymbol{X}_i'\boldsymbol{\beta})\right]^{(1-Y_i)} \tag{3.5}$$

Similarly, the likelihood function to get MLE of $\boldsymbol{\beta}$ with the logit model is:

$$L(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X}) = \prod_{i=1}^{N} \left[\frac{exp(\boldsymbol{X}_i'\boldsymbol{\beta})}{1+exp(\boldsymbol{X}_i'\boldsymbol{\beta})}\right]^{Y_i} \left[\frac{1}{1+exp(\boldsymbol{X}_i'\boldsymbol{\beta})}\right]^{1-Y_i}. \tag{3.6}$$

To make 3.5 and 3.6 easier to work with, we take logarithm on the probit and logit likelihood because if $\boldsymbol{\beta}$ maximize $L(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X})$, then it also maximize $logL(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X})$, which is concave and the estimates $\hat{\boldsymbol{\beta}}$ are the points at which the derivative equals to 0. The MLE $\hat{\boldsymbol{\beta}}$ is the solution to the equations: $\frac{\partial logL(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X})}{\partial \boldsymbol{\beta}} = 0$. Neither of these solutions have a closed

form. Instead, iterative algorithms such as EM (Dempster et al., 1977) or Newton-Raphson algorithm must be used.

Iteratively Reweighted Least Squares (IRLS) is also a popular algorithm to find the MLE. For a logit model, let $E(Y_i) = \frac{e^{\boldsymbol{X}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i\boldsymbol{\beta}}}$ and $Var(Y_i) = V_0$, $V = Diag(V_0, \ldots, V_0)$, $i = 1, \ldots, m$. If $\boldsymbol{\beta}$ is known, a simple consistent estimator of V would be

$$\hat{\boldsymbol{V}}_0 = \frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \frac{e^{\boldsymbol{X}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i\boldsymbol{\beta}}}\right)\left(Y_i - \frac{e^{\boldsymbol{X}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i\boldsymbol{\beta}}}\right)'.$$

The algorithm is: start with the OLS estimator and compute $\hat{V}$ with OLS of $\boldsymbol{\beta}$, then replace V by $\hat{V}$ just obtained to get the next step estimator of $\boldsymbol{\beta}$, and repeat the process. Goldstein (1989) showed that if normality holds and the IRLS converges, the limiting estimator is identical to the Maximum Likelihood Estimator. Hence, Iteratively Reweighted Least Squares is widely used to find both the maximum likelihood estimates and the asymptotic covariance matrix for parameters in generalized linear models (which include the probit and logit models). For more detail about IRLS, see Holland and Welsch (1977) and Green (1984).

For the large samples, MLEs exhibit the asymptotic properties of unbiasness, efficiency, and normality. These properties seem to hold reasonably well for probit and logit models, even in moderate-sized samples in which number of observations (N) is only 100 more than the number of predictor variables (p)($N - p = 100$). However, the maximum likelihood method works poorly when the sample size is small, and the MLE can be substantially biased when the samples are small (Griffiths and Pope, 1987). Another drawback to MLEs with these models is that since the likelihood equations for probit and logit are nonlinear, the algebraic solutions are not obtainable, and the approximations by standard iterative algorithm need to be used.

In summary, as long as the assumptions hold, the MLE estimators for probit and logit models have nearly the same asymptotic properties as OLS estimates of the linear regression model. When the MLE models are nonlinear, computational cost can increase, and the properties are asymptotic. It is known, under regularity conditions, that MLEs are asymp-

totically normal with asymptotic covariance matrix equal to the Fisher Information matrix. One representation of the (i, j)th element Fisher Information matrix of $\hat{\boldsymbol{\beta}}$ is:

$$-E\left[\frac{\partial^2 logL(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X})}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j}\right],$$

which can be estimated by the iterative maximization algorithm such as Newton-Raphson or IRLS algorithm and is obtainable from most computer packages. The Fisher Information matrix can be used to construct asymptotic confidence intervals and hypothesis tests.

Bayesian Estimation for Probit and Logit Regression Models

An alternative to classical estimation (Least Square, Maximum Likelihood, etc) is to use a Bayesian method of inference. In general, the MLE $\theta_{ML} = argmax_\theta L(\theta|X)$. If we interpret the ML estimation from the Bayesians view, the MLE can be seen as a Bayesian estimator assuming uniform priors $\pi(\theta) \sim 1$. The MLE is the posterior mode. However, assuming an uniform prior is often unreasonable, since we frequently have information about the distribution of the parameter we are estimating, and uniform priors may also lead to improper posteriors.

The Bayesian paradigm for statistical inference is revolves around Bayes Theorem. Bayesian can trace their roots back to Bayes (1763) and Laplace (1813). The fundamental idea of Bayesian statistics is that we have unobserved parameters, $\boldsymbol{\theta}$, which follows a prior distribution, $\pi(\boldsymbol{\theta})$, the observed data $\boldsymbol{Y}$ are conditional on the value of $\boldsymbol{\theta}$, are assumed to have a distribution $f(\boldsymbol{Y}|\boldsymbol{\theta})$, which is the likelihood. MLE bases the inferences on $\boldsymbol{\theta}$ primarily on the likelihood. In contrast, Bayesian inference is centered around the posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$, the conditional distribution of unobserved parameter given the observed data. The posterior distribution is calculated using Bayes' theorem and provides a natural updating of prior belief based on what was observed in the data. The posterior distribution is formulated

as:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{Y}) &= \frac{f(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\boldsymbol{Y})} \\ &\propto f(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),\end{aligned} \tag{3.7}$$

where, $m(\boldsymbol{Y}) = \int f(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal likelihood of the data. The prior distribution represents the information about an uncertain parameter $\boldsymbol{\theta}$ that is combined with the probability distribution of new data to yield the posterior distribution. There are generally two ways of defining a prior distribution: subjective and objective. The subjective approach holds a belief that the prior distribution should reflect a researcher's or expert's actual subjective opinion about the unobserved parameters, and the opinion is incorporated into the posterior distribution of the parameter. While objective Bayesian minimize the influence of prior distribution on the posterior, meaning that the observed data dominate the posterior distribution. The main types of objective priors are uniform priors, Jeffrey prior or reference priors.

Albert and Chib (1993) developed a fully Bayesian approach to estimate $\boldsymbol{\beta}$ in the probit model. They used a latent variable S_i, which is continuous and unknown, but conditional on the observed binary variable Y_i:

$$Y_i = \begin{cases} 1 & \text{if } S_i > 0, \\ 0 & \text{if } S_i \leqslant 0, \end{cases}$$

and $\boldsymbol{S}_i$ is modeled as:

$$S_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \tag{3.8}$$

with

$$\epsilon_i \overset{iid}{\sim} N(0,1),$$

for $i = 1, 2, \ldots, N$. It is easy to show that:

$$P(Y_i = 1|\boldsymbol{X}_i\boldsymbol{\beta}) = P(S_i > 0) = \Phi(\boldsymbol{X}_i'\boldsymbol{\beta}).$$

Conditional on Y_i, S_i follows a truncated normal distribution $S_i|\boldsymbol{\beta}, Y_i \sim N(\boldsymbol{X}_i'\boldsymbol{\beta}, 1)$, truncated to the possible positive or negative range of S_i indicated by each Y_i. The joint posterior density of the latent variable $\boldsymbol{S} = (S_1, S_2, ..., S_N)$ and coefficient $\boldsymbol{\beta}$ given the data $\boldsymbol{Y} = (Y_1, ..., Y_N)$ is:

$$\pi(\boldsymbol{\beta}, \boldsymbol{S}|\boldsymbol{Y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \{I(S_i > 0)I(Y_i = 1) + I(S_i \leq 0)I(Y_i = 0)\} \times \phi(S_i; \boldsymbol{X}_i'\boldsymbol{\beta}, 1),$$

where $\phi(X; \mu, \sigma^2)$ is the normal pdf $N(\mu, \sigma^2)$. $I(X \in A)$ is the indicator function that equals to 1 if the random variable is contained in the set A. $\pi(\boldsymbol{\beta})$ is the prior for $\boldsymbol{\beta}$. This joint posterior distribution is not of closed form, making it difficult to normalize and construct inference directly.

Using a flat (noninformative) prior for $\boldsymbol{\beta}$, the conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{S}$ and $\boldsymbol{Y}$ is:

$$\boldsymbol{\beta}|\boldsymbol{S}, \boldsymbol{Y} \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \phi(\boldsymbol{S}; \boldsymbol{X}'\boldsymbol{\beta}, 1)$$

$$\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{S} \sim N\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{S}, \quad (\boldsymbol{X}'\boldsymbol{X})^{-1}\right).$$

$\boldsymbol{S}_i|Y_i, \boldsymbol{\beta}$ is truncated normal $N(\boldsymbol{X}'\boldsymbol{\beta})$, which truncated at the left by 0 if $Y_i = 1$, and truncated at the right by 0 if $Y_i = 0$.

Albert and Chib (1993) described a Gibbs sampler for the probit model. Introducing the latent variable makes the the conditional distributions of the model parameters in the probit model conjugate, if the flat prior or a proper normal prior for $\boldsymbol{\beta}$ were assumed, the full conditional distribution of all parameters can be easily calculated. They used Gibbs sampling to sample from the full conditional distributions. See Section 3.2.2 for a review of Gibbs sampling. Since the esoteric interpretation of the probit models is widely criticized, a Bayesian version logit model for binary or polychotomous response data was proposed by O'Brien and Dunson (2004). O'Brien and Dunson (2004) introduced a fully Bayesian logit model using a similar latent continuous variable, S_i, behind the observed binary or ordered categorical data. In their model, S_i is an independent logistically distributed random variable

with location $\boldsymbol{X}'\beta$ and scale $\sigma^2_{\epsilon_{logistic}} = 1$. The density function of S_i they used is:

$$L(S_i|\boldsymbol{X}'\boldsymbol{\beta}) = \frac{exp\{-(S_i - \boldsymbol{X}'\boldsymbol{\beta})\}}{[1 + exp\{-(S_i - \boldsymbol{X}'\boldsymbol{\beta})\}]^2}$$

Or expressed as:

$$S_i = \boldsymbol{X}\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} logistic(0, 1)$$

The logistic distribution has similar tail behavior as the t-distribution, so O'Brien and Dunson (2004) obtained the posterior multivariate logistic distribution from multivariate t distribution by importance sampling. Due to close correspondence between logistic and t distribution, selecting the degree of freedom in the t distribution $v = 7.3$ and setting the parameter $= \pi^2(v-2)/3v$, the t distribution can approximate the logistic distribution. Since the posterior computation is challenging, caused by the complexity of the likelihood, they sampled from an alternative t-link model and used importance sampling to estimate the parameters. See Section 3.2.2 for more details on importance sampling.

Related work in Bayesian logit models, such as Jaakkola and Jordan (2000) used the variational estimation techniques to approximate the posterior, and the EM algorithm to estimate the posterior mode for Bayesian logistic models with Gaussian prior distribution. Xu and Akella (2008) used a Bayesian logistic regression model for active relevance feedback analysis. These belongs to the empirical Bayes and modeling strategy is not a fully Bayesian approach. An example of a fully Bayesian model is Genkin et al. (2007), who presented a simple Bayesian logistic regression approach that uses a Laplace prior to avoid overfitting and produces sparse predictive models. Holmes and Held (2006) discussed auxiliary variable methods for inference in Bayesian logistic regression, including covariate set uncertainty, and showed how the logistic method is easily extended to multinomial regression models.

3.1.2 Multilevel Linear, Probit and Logit Models

In this section, we will review multilevel linear, probit and logit models.

Multilevel/Hierarchical Linear Models

Multilevel linear models are also called hierarchical linear models (Goldstein, 1995; Wong and Mason, 1985), mixed effect models and random-effects models (Singer, 1998), or random-coefficient regression models (Longford, 1993), and covariance component models (Dempster et al., 1981). Raudenbush et al. (2002) classified them all "Hierarchical Linear Models". In social and behavioral sciences, these multilevel analysis techniques has gained increasingly popularity, and many applications have been published.

Multilevel (hierarchical) linear models have become popular in education research for more than two decades. For example, Aitkin and Longford (1986) used what they called "random parameter models" for the analysis of the studies involving clustered observations to access school effectiveness. Lockheed et al. (1989) applied a multilevel model to analyze what improved performance in grade 8 mathematics in Thailand. Goldstein (1989) developed an iterative generalized least squares estimation procedure for analyzing hierarchically structured data for an example in education data, and made an efficient computational procedures later in 1992. A complete book "Multilevel Statistical Models" by Goldstein was published in 2011. Statistical computing programs were also created for computing such as HLM (Raudenbush et al., 2004), MIXOR (Hedeker and Gibbons, 1996), MLWIN (Rasbash et al., 2000), SAS $PROC\ MIXED$ procedure (Littell, 2006), and VARCL (Longford, 1988).

The name of multilevel/hierarchical models is from the data structure: first we have variables describing individuals, but individuals are also grouped in larger units, each unit consisting of a number of individuals, and these units may be grouped in larger units. Longitudinal data, which records individuals responses over time while the responses at the same time are correlated with each other, can also use a multilevel structure. For example, students (the first level) are grouped in classes, and classes (the second level) are grouped in schools, schools (the third level) in school districts, and so on. In other words, data are nested or clustered into different levels of groups. The advantages of using multilevel/hierarchical

models if the data is grouped or nested into different levels are: First, individuals in the same group have same value of the group variable, and have some similarities. For example, students' grades in an Intermediate Algebra class can be different than the grades in a dancing class. So the assumption of independence of observations, that is basic for classical statistical techniques, does not apply here. Second, with multilevel/hierarchical models, we take into account the variation within groups and between groups, or variation between individuals within the same group and the variation between groups at each level. Third, multilevel models can provide good information about group effect at different levels. For example, using individual response variable about students, we get information about the variability of students in the response between the upper level groups like classes, even the variability between schools (the third level group). Regular one-level models assume all the variability at the individual response level and do not give information about group effects, this can be misleading (Bosker and Snijders, 1999). Hierarchical Linear Models have become the basis of two of the most active research areas in statistics and biostatistics: nonlinear mixed models and generalized linear mixed models.

The linear regression model was expressed in Equation 3.1. Hierarchical Linear Models are statistical models in the analysis of variance (ANOVA) and in regression analysis where it is assumed that some of the coefficients are fixed while others are random. A standard two level Hierarchical Linear model takes the form:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{\tau} + \boldsymbol{\epsilon}_i. \tag{3.9}$$

With,

$$\begin{aligned} \text{The first level} : \epsilon_i &\overset{iid}{\sim} N(0, \sigma_\epsilon^2), \\ \text{The second level} : \tau_j &\overset{iid}{\sim} N(0, \sigma_\tau^2), \end{aligned}$$

where $\boldsymbol{X}_i = (1, X_{i1}, \ldots, X_{ip})$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$, and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_j)$, $\boldsymbol{Z}_i$ is a vector that contains only 0s and 1s, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, J$. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the fixed and random design vectors, respectively. $\boldsymbol{\beta}$ is a vector of unknown fixed effects, and $\boldsymbol{\tau}$

is the vector of unknown random effects. The response variable at level 1 is assumed to be normally distributed when the outcome variable is continuous. Another assumption for multilevel linear models or hierarchical linear models is that the expected outcome at each level may be represented as a linear function of the regression coefficients (Raudenbush et al., 2002).

Bayesian models naturally incorporate random effects since all parameters are assumed to be random variables. To make a hierarchical linear model a Bayesian model, we only need to assume a prior for the nonrandom parameters. We discuss Bayesian computation in Section 3.2.2.

Multilevel Probit and Logit Models

Multilevel logistic regression models add random effects into the ordinary (one-level) fixed effect logistic regression model for the data those are nested or clustered into different levels. Thus, similar to multilevel linear models, with $P_i = P(Y_i = 1)$, the multilevel logit model has both fixed-effects and random effects and is in the form:

$$logit(P_i) = log(\frac{P_i}{1-P_i}) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau} \Longrightarrow P_i = \frac{exp(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau})}{1 + exp(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau})}. \tag{3.10}$$

Alternatively, a multilevel probit model is formulated as:

$$\Phi^{-1}(P_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau} \Longrightarrow P_i = \Phi(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau}) = P(Z \leq \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau}). \tag{3.11}$$

where Z has standard normal distribution.

Non-Bayesian Estimation Methods for Multilevel Linear, Probit and Logit Models

Standard non-Bayesian methods of estimation in hierarchical generalized linear models are Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML). Hartley and Rao

(1967) first used ML in mixed model analysis. They developed a procedure for the maximum likelihood estimation of the unknown constants and variances included in the general mixed analysis of variance model, involving fixed and random factors and interactions, also discussed the consistency and asymptotic efficiency of the estimates as well as tests of hypotheses and confidence regions. For a detailed description of ML or REML estimation for multilevel linear models, see Jiang (2007) and Raudenbush et al. (2002).

The REML method was first proposed by Thompson (1962) and was put on a broad basis by Patterson and Thompson (1971). It was to take care of the bias of the MLE for the variance components when the number of parameters increase with the sample size. A good example is the well-known Neyman-Scott problem (Neyman and Scott, 1948). The model for this problem can be expressed as: $y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, \ldots, m, j = 1, \ldots, n$, where ϵ_{ij} are independent and distributed as $N(0, \sigma^2)$, and we are interested in estimating σ^2. It can seen as a special case of the hierarchical linear model when $n = 2$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\tau} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is a vector of observations, $\boldsymbol{X}$ is a matrix of known covariates, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. $\boldsymbol{Z}$ is a known design matrix, $\boldsymbol{\tau}$ is a vector of random effects, and $\boldsymbol{\epsilon}$ is a vector of errors. It can shown in this case that the MLE of σ^2 is not consistent as $m \to \infty$. Here, the number of parameters is proportional to the sample size, while the means $\mu_1, \ldots, \mu_m$ are considered as nuisance parameters. Instead of using the original data, we do a simple transformation: $z_i = y_{i1} - y_{i2}$, so the nuisance terms are gone, and $z_i \sim N(0, 2\sigma^2)$. The key idea of REML is: apply a transformation to the data to eliminate the fixed effects, then use the transformed data to estimate the variance components. Once the REML estimator of the variance components is obtained, the MLE of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ can be calculated. There are other solutions to the Neyman-Scott problem, see Berger and Bernardo (1992) for a Bayesian solution.

The REML estimator does not depend on the choice of transformation. Under suitable conditions, the REML estimator is consistent and asymptotically normal. Jiang (1996, 1997)

showed that if the rank of the design matrix $\boldsymbol{X}$ is fixed or bounded, the ML and REML estimators are asymptotically equivalent, and REML is obviously superior to ML where the number of fixed effects grows with the sample size.

As an unique case of generalized hierarchical linear models, multilevel logit or probit models have the nonlinear first level model, and the first level error terms are non-normal. The likelihood functions for $\boldsymbol{\beta}$ under a generalized hierarchical linear mixed model are not of closed form. In fact, these likelihood functions involve high-dimensional integrals that can not be evaluated analytically. Using an example from Jiang (2007), for the simple one-way random effect model with binary data y_{ij}, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$ are conditionally independent, with $P_{ij} = P(y_{ij} = 1|u, v)$, $logit(P_{ij}) = u + u_i + v_j$, u is an unknown parameter. The random effects $u_1, \ldots, u_{m1}$ and $v_1, \ldots, v_{m2}$ are independent such that $u_i \sim N(0, \sigma_1^2)$, $v_j \sim N(0, \sigma_2^2)$, where σ_1^2 and σ_2^2 are unknown. It can be shown that the likelihood function for estimating $(u, \sigma_1^2, \sigma_2^2)$ can be expressed as:

$$c - \frac{m_1}{2} log(\sigma_1^2) - \frac{m_2}{2} log(\sigma_2^2) + uy_{..} + log \int \ldots \int \left[\prod_{i=1}^{m_1} \prod_{j=1}^{m_2} (1 + exp(u + u_i + v_j)) \right]^{-1}$$
$$\times exp \left[\sum_{i=1}^{m1} u_i y_{i.} + \sum_{j=1}^{m2} v_j y_{.j} - \frac{1}{2\sigma_1^2} \sum_{i=1}^{m_1} u_i^2 - \frac{1}{2\sigma_2^2} \sum_{j=1}^{m_2} v_j^2 \right] du_1 \ldots du_{m1} dv_1 \ldots dv_{m_2}.$$

The high dimensional integral in the above equation is very difficult to evaluate, if $m = n = 40$, the dimension of the integral will be 80, and the integrand involves a product of 1600 terms. This shows that for multilevel probit and logit models, the likelihood-based inference can be computationally challenging in many situations.

To solve or avoid this difficulty, some numerical integration or approximation approaches must be employed. Stiratelli et al. (1984) estimated the parameters of logistic regression model with nested, normally distributed random effects by approximating the mode of the posterior distribution of the random parameters. Inference is based upon maximum likelihood estimation of fixed effects and variance components, as well as a empirical Bayes estimation of random effects. However, this method is computationally intensive. Wong

and Mason (1985) used essentially the same approach, and it has been extended by many authors. Breslow and Clayton (1993) termed this method penalized quasi-likelihood (PQL). Gilmour et al. (1985) considered general covariates in logistic regression with a Gaussian random intercept using expectation maximization (EM) and Newton-Raphson algorithms, respectively. Anderson and Aitkin (1985) first proposed a random effect binary model with maximum likelihood estimator of variance component, an univariate Gaussian quadrature was used for integration of the random effects.

Many other methods were developed for the approximation of the high dimensional integrals in Maximum Likelihood Estimation such as Laplace approximations, the Newton-Raphson algorithm, normal approximations, and Gauss-Hermite quadrature. We leave these methods to Section 3.2.1. Most of the standard software packages have procedures for these methods.

Bayesian Multilevel Probit and Logit Regression Models

The computational burden has limited data analysis with multilevel probit and logit models with which the outcome variable $\boldsymbol{Y}$ is dichotomous rather than Gaussian in several ways.

1. First, investigators have largely restricted their attention to random intercept models to avoid higher dimensional numerical integration. Because the likelihood methods that work with Gaussian outcomes do not work in this case, the likelihood function itself cannot even be evaluated without integrating out the random effects.

2. Second, specialized software is required and is typically optimized for a particular random effects distribution (e.g., the Gaussian). Available software such as SAS, R, MLwiN fit nonlinear models via quasi-likelihood methods by linearizing the second line of the model via Taylor series expansion, yielding marginal and penalized quasi-likelihood estimates according to the form of the expansion used. These are not full likelihood methods and would be better termed likelihood-based techniques. Draper

(2008) showed nominal 95% interval estimates with this approach in random effects logistic regression models can be far less than 95% when the intervals are based only on marginal and penalized quasi-likelihood point estimates and their (estimated asymptotic) standard errors.

3. Third, inferences about the regression coefficients have often been made conditional upon the estimated random effects variance to avoid difficult integration.

To overcome these computational difficulties, Zeger and Karim (1991) first adopted a Monte Carlo method, the Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984), and casted Generalized Hierarchical Linear Model or Linear Mixed Models (GLMM) in the Bayesian framework to overcome the computational limitations with maximum likelihood estimation. They provided a fully Bayesian approach to the multilevel/hierarchical linear model: $y_{ij} = x'_{ij}\beta + z'_{ij}\tau_i + \epsilon_{ij},\ i = 1, \ldots, m, j = 1, \ldots, n.$ Where y_{ij} is the response for the jth observation in cluster i and here y_{ij} follows an exponential family distribution; x_{ij} is a p x 1 vector of covariates associated with that response; β is the vector of regression coefficients, and z_{ij} is a q x 1 subset of x_{ij} with random effects; τ_i is a q x 1 vector of random effects assumed to be normally distribution with mean 0 and variance σ^2_τ, and ϵ_{ij} is independent error with mean 0 and variance ϵ^2_ϵ. Compared to a GLM, this multilevel model reflects heterogeneity across clusters in the regression coefficients, since observations from the same cluster are correlated. Instead of maximum likelihood estimation, Markov Chain Monte Carlo rejection sampling methods (Devroye, 1986) and Gibbs sampling were used to estimate the parameters.

Zeger and Karim (1991) first implemented Gibbs sampling and rejection sampling to generalized linear mixed models with Gaussian random effects, and estimated the generalized linear mixed models with a fully Bayesian approach. Albert and Chib (1993) brought a fully Bayesian version of probit model on binary and polychotomous response data, and used the Gibbs sampling in conjunction with data augmentation, which lead to computationally simpler strategy. Bradlow et al. (1999) added a random effect to the standard Item Response

Theory (IRT) models to account for the nesting of items within the same testlets. They adopted a latent variable probit model. In their model, the latent score t_{ij} is given by:

$$t_{ij} = a_j(\theta_i - b_j) + \epsilon_{ij},$$

where,

$$y_{ij} = \begin{cases} 1 & \text{if } t_{ij} > 0, \\ 0 & \text{if } t_{ij} \leqslant 0, \end{cases}$$

and ϵ_{ij} is a unit normal variance used to indicate randomness in response y_{ij} across hypothetical replication of item i. The parameter θ, a_j and b_j were given interpretation as examinee proficiency, item discrimination, and item difficulty, respectively. Independent normal priors were given to each parameter:

$$\begin{aligned} \theta_i &\sim N(0, 1), \\ a_j &\sim N(u_a, \sigma_a^2), \\ b_j &\sim N(u_b, \sigma_b^2), \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2). \end{aligned}$$

The conditional distribution of the latent score t_{ij} given other parameters is truncated normal, and the conjugate Gaussian priors yield Gaussian conditional distributions for each parameter, and similar to Albert and Chib (1993), a Gibbs sampler was used. Modifications of regular multilevel probit and logit models were presented with different priors. For example, Kleinman and Ibrahim (1998) proposed a Dirichlet process prior for the distribution of the random effects. Fahrmeir and Lang (2001) developed a unified approach for Bayesian inference via Markov Chain Monte Carlo simulation in generalized additive and semiparametric mixed models, by assigning appropriate Markov random field priors with different forms and degrees of smoothness.

3.1.3 Ordered and Multinomial Probit/Logit Models and Estimation

In this section, we review the ordered probit/logit models, and multinomial probit/logit models, as well as the estimation approaches to these models.

Ordered and Multinomial Probit/Logit Models

Both derived from binary probit/logit models, ordered probit/logit models fit ordinal categorical data, while multinomial probit/logit models work for unordered nominal data.

Ordered Probit/Logit Models Ordinal data are the most frequently encountered type of data in the social sciences. For example, respondents are asked to characterize their opinions on a scale such as: "yes, maybe, no", "always, frequently, sometimes, rarely, never", or a Likert scale: "strongly Disagree, disagree, don't know, agree, strongly agree" or "strongly dissatisfied, dissatisfied, neutral, satisfied, dissatisfied". These categories have a clear order. We also run into data like: "no high school diploma, high school diploma, some college, bachelor's degree, masters degree, doctoral degree", "free school lunch, reduced school lunch, full price lunch", "0-10K per year, 10-20K per year, 20-30K per year, 30-60K per year, $>$ 60K per year", "low, medium, high", "basic math, regular math, pre-AP math, AP math". The similarity of these data is that there is a increasing or decreasing order on the data according to the strength or amount, and there commonly assumed to be equal distance between the neighboring points. The ordered logit and probit models enable us to model ordinally scaled dependent variables with one or more independent variables. The most natural way to view ordinal data is to postulate an underlying latent variable S_i associated with each observed

response Y_i. For binary cases, the response data has only two categories, 0 or 1.

$$Y_i = \begin{cases} 1 & \text{if } S_i \geq \gamma, \\ 0 & \text{if } S_i < \gamma. \end{cases}$$

Where the value of γ is determined by the specific situation in application.

If the response data has many ordered categories (Albert and Chib, 1993):

$$Y_i = \begin{cases} 1 & \text{if } \gamma_0 < S_i < \gamma_1; \\ 2 & \text{if } \gamma_1 < S_i < \gamma_2; \\ 3 & \text{if } \gamma_2 < S_i < \gamma_3; \\ 4 & \text{if } \gamma_3 < S_i < \gamma_4; \\ ... & \\ m & \text{if } \gamma_c - 1 < S_i < \gamma_c. \end{cases}$$

Where $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_c)$ are the boundaries of the latent variable for each category. The probability of an observation falling in each category is:

$$Pr(Y_i = c) = Pr(\gamma_{c-1} < S_i < \gamma_c) \tag{3.12}$$

and the value of bounds for the latent variable S_i corresponding to each category is decided according to each specific situation. The parameters $\boldsymbol{\gamma}$ are also called *thresholds* or *cutpoints*, and are in increasing order $(-\infty < \gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_{c-1} < \gamma_c < \infty)$. Such latent variables are assumed to be continuous and centered on a mean value, which are often modeled as a linear function of one or several covariates. The distribution of the latent variable can be expressed with Equation 3.13

$$\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.13}$$

where the covariance matrix $cov(\boldsymbol{\epsilon}) = \Sigma$. Then the probability of Y_i falling in each category

is:

$$\begin{aligned} P_{ic} &= \int_{\gamma_{c-1}}^{\gamma_c} f(S_i - \boldsymbol{x}_i'\beta) dS_i, \\ &= P(\gamma_{c-1} < S_i < \gamma_c), \\ &= F(\gamma_c - \boldsymbol{X}_i'\beta) - F(\gamma_{c-1} - \boldsymbol{X}_i'\beta). \end{aligned} \tag{3.14}$$

Where $F(\cdot)$ is the cumulative distribution function (CDF).

Identifiability in the latent probit model is always a concern, because if there is an existence of unidentifiable parameters, trajectories of the Markov chain for components of the parameter will tend to drift to a very extreme value, are difficult to converge, and lead to unstable computation (Gelfand and Sahu, 1999). In the ordered probit model, the parameters $\boldsymbol{\beta}$ and the covariance matrix Σ are not identifiable if the boundaries of $\boldsymbol{\gamma}$ are unknown, see McCulloch et al. (2000). To ensure the identifiability of the parameters, we either have to impose known boundaries, or at least restrictions, on the latent variable. For example, make $\gamma_1 = 0$, or fix or restrict the covariance matrix. Albert and Chib (1993) made the covariance matrix equal to I. Let σ_{ij} be the ijth element of the covariance matrix Σ, since σ_{11} is positive, frequentist methods often set $\sigma_{11} = 1$ to achieve the identification of parameters. McCulloch et al. (2000) developed a Bayesian algorithm to leave the boundaries of the latent variable random, but work on priors of the covariance matrix to make the parameters in the multinomial probit model identifiable.

Let $\boldsymbol{Y}_i$ be an ordinal response variable with C categories for ith subject, along with a vector of covariates $\boldsymbol{X}_i$. A regression model can be done to establish the relationship between the response variable and the covariates, and yield the probability that a individual falling in each category $p_{ci} = P(\boldsymbol{Y}_i = y_c|\boldsymbol{X}_i)$, $c = 1, \ldots, C$. Ordered probit or logit models can be done with SAS procedures such as *proc logistic*, *proc genmod* and R *polr* or *lrm* procedures, or SPSS and other software. Usually, these models give the cumulative probabilities $g_{ci} = P(\boldsymbol{Y}_i \leq y_c|\boldsymbol{X}_i)$, $c = 1, \ldots, C$. The last cumulative probability equals 1. Thus the probability that an individual's response falls in category c can be obtained by using the cumulative probability of the category minus the probability of the category before it: $P(\boldsymbol{Y}_i = c) =$

$P(\boldsymbol{Y}_i \leq c) - P(\boldsymbol{Y}_i \leq c-1)$. An ordered logit model for an ordinal response $\boldsymbol{Y}_i$ with C categories is defined by a set of $C-1$ equations where the cumulative probabilities $p_{ci} = P(\boldsymbol{Y}_i \leq y_c|\boldsymbol{X}_i)$ are related to the predictors $\boldsymbol{X}_i$ through the function:

$$logit(p_{ci}) = logit(\frac{p_{ci}}{1-p_{ci}}) = \gamma_c - \boldsymbol{X}_i'\boldsymbol{\beta},\ c = 1, \ldots, C.$$

It is not possible to simultaneously estimate the overall intercept β_0 and all $C-1$ thresholds, and the identification problem is usually solved by either omitting the overall constant from the linear predictor (to make β_0 =0) or fixing the first threshold to zero (make γ_1 =0). Where

$$p_{ci} = \frac{exp(\gamma_c - \boldsymbol{X}_i'\boldsymbol{\beta})}{1 + exp(\gamma_c - \boldsymbol{X}_i'\boldsymbol{\beta})} = \frac{1}{1 + exp(-\gamma_c - \boldsymbol{X}_i'\boldsymbol{\beta})}.$$

The ordered logit model is also called the "proportional odds model" because the ratio of the odds for the event $Y_1 \leq c$ to the odds of the event $Y_2 \leq c$ is:

$$\frac{p_{c1}/(1-p_{c1})}{p_{c2}/(1-p_{c2})} = exp\left[-(\boldsymbol{X}_1 - \boldsymbol{X}_2)'\boldsymbol{\beta}\right],$$

which is independent of the category of response c. The ordered logit model is a member of the wider class of *cumulative ordinal models*, where the logit function is replaced by other link function, a probit or complementary log-log are the most common ones. In the case of a probit link, the probability of an observation being in category c is:

$$\Phi^{-1}(p_{ic}) = \gamma_c - \boldsymbol{X}_i'\boldsymbol{\beta}_c$$

$$\begin{aligned} Pr(Y_i = c|\boldsymbol{X_i}, \boldsymbol{\beta}_c, \sigma_\epsilon^2) &= Pr(S_i \geq \gamma_c|\boldsymbol{X}_i, \boldsymbol{\beta}_c, \sigma_\epsilon^2) \\ &= 1 - \Phi(\gamma_c - \boldsymbol{X}_i\boldsymbol{\beta}_c) \\ &= \int_{\boldsymbol{X}\beta}^{\infty} \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}S_i^2\right] dS_i. \end{aligned}$$

This is also called *ordinal probit model*. The ordinal probit model predicts probabilities similar to those obtained from the proportional odds model, just as predictions from binary probit model are similar to the obtained from a binary logit model. Since ordinal probit model can make sampling from its posterior distribution particularly efficient, for this reason, it may be preferred to other model links if a Bayesian model is employed.

Unordered Multinomial Probit/Logit Models Unordered multinomial probit/logit models apply to the data which are mutually exclusive categories with no logical order, for example, marital status: "married, divorced, widowed, never married", political parties: "Republican, Democrat, Green", or occupational choice: "academic, business, non-profit organization". Unordered multinomial probit/logit models are also a generalization of binary probit/logit models. For the unordered multinomial logit model, the probability of an individual i choose alternative j is:

$$P_{ij} = P(Y_i = j) = \frac{exp(X_i'\boldsymbol{\beta}_j)}{\sum_{j=1}^{J} exp(X_i'\boldsymbol{\beta}_j)}, \tag{3.15}$$

where $\boldsymbol{X}_i$ represents the independent variables for individual i, $\boldsymbol{\beta}_j$ is the coefficient of the independent variables for category j, and $\sum_{j=1}^{J} p_{ij} = 1$. To estimate the model, we set one category as baseline (reference), and $\boldsymbol{\beta}_1 = 0$. So when there are only two categories, the denominator becomes $1 + exp(\boldsymbol{X}_i'\boldsymbol{\beta}_j)$, and Equation 3.15 becomes the binary logit model.

Similar to the difference of the binary logit model to the binary probit model, a multinomial probit model uses the standard normal CDF as the link function. One important difference between the multinomial logit model and multinomial probit model is that the logit model assumes independence of irrelevant alternatives (IIA), while the probit model does not. IIA requires that an individual's evaluation of an alternative relative to another alternative should not change if more alternatives are added or dropped to the analysis. A common example of this assumption is the red bus-blue bus problem. Suppose we have three transportation categories: a red bus, a blue bus, and a car, and we are equally likely to choose each one. Thus the probability of selecting each is 0.33, but if a red bus is removed as an option, then we will be twice as likely to take the blue bus than before, and the probability of taking the blue bus and the car becomes 0.66 and 0.33, because the red bus and blue bus are perfect substitute of each other. Multinomial logit regression assumes that none of the categories can serve as substitutes. This is often erroneous. The unrealistic assumption of IIA can be relaxed in the ordered logit model by introducing a latent variable into the model according to the order indicated by the response data.

Parameter estimation of the multinomial probit or logit models is the same as estimating a series of binary logit or probit models. To estimate the model, one category has to be chosen to be the baseline (reference) category for the other categories, so the coefficient of independent variables for this category need to be normalized to 0, and the estimation is made for each of the other categories versus the baseline category. Thus the coefficient interpretation for category j is: in comparison to the baseline category, each unit increase in the independent variable increase/decrease the probability of selecting of alternative j, by the marginal effect expressed as a percent. See So and Kuhfeld (1995) and Chikhi (2013) for a detailed description of multinomial logit models.

Multilevel Ordered Probit/Logit Models Multilevel (random effect) ordered probit and logit models can be used for the analysis of correlated or nested ordinal response data. Rampichini and d'Andrea (1998) proposed a hierarchical ordinal probit model with group structure and ordinal response variables to predict the life satisfaction in Italy in different regions. A latent variable was also used in their model. Estimation of the model was obtained by means of an iterative maximum likelihood estimation procedure based on numerical integration and the EM algorithm. Fielding et al. (2003) developed a multilevel logit model to analyze students' grade data in different institutions. They treated the students' grades as ordinal data, contrasted to converting the letter grades to point score. Item response models can be considered as latent variable models that have an individual-specific underlying trait that drives a respondent's multiple responses to a set of questions or items. See Section 3.1.4 for an introduction of item response models.

Adams et al. (1997) proposed two-level multilevel item response models to model categorical response data with a latent variable. They used a vector valued random variable $\boldsymbol{X}$ to indicate the $K_i + 1$ possible responses to item i. That is, $\boldsymbol{X}' = (X_{il}, X_{i2}, \ldots, X_{iK_i})$, where

$$X_{ij} = \begin{cases} 1 & \text{if response to item i is in categor } j; \\ 0 & \text{otherwise.} \end{cases}$$

For the sake of model identification, a response in category zero is denoted by a vector of zeros as a reference category. They fit a logit link to the model, and estimated the model with maximum likelihood using the Newton-Raphson method or the EM algorithm (Dempster et al., 1977), efficient posterior computation with the model is challenging, due to the lack of simple forms for the posterior distributions.

The decision between linear regression and ordered multinomial regression is not always black and white. When you have a large number of categories that can be considered equally spaced, simple linear regression is an optional alternative (Gelman and Hill, 2007), but the assumptions of a simple linear regression must be satisfied.

Maximum Likelihood Estimation to Ordered Probit/Logit Models

Maximum likelihood estimation for ordinal regression models may be obtained using iterative reweighted least squares (IRLS), an estimate of the asymptotic covariance matrix of the MLE can be obtained with this method to perform classical inference concerning the MLE. If an uniform prior is assumed in a Bayesian model, then the MLE and asymptotic covariance matrix obtained using IRLS can be used as approximation to the posterior mode and posterior covariance marix.

For non-Bayesian ordered probit models, Rampichini and d'Andrea (1998) proposed a hierarchical ordinal probit model with group structure and ordinal response variables to predict the life satisfaction in Italy in different regions. Their latent variable multilevel probit model can be written as:

$$\begin{aligned} S_{in} &= \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \tau_j + \epsilon_{ij}, \quad i = 1, \ldots, K, n = 1, \ldots, N_i; \\ Y_{inj} &= I(\gamma_{j-1} < Y^*_{in} \leq \gamma_j), \quad j = 1, \ldots, m \end{aligned}$$

where S_{in} is the latent variable corresponding the categories Y. $\boldsymbol{X}'_{in}$ is a vector of observable explanatory variables, $\boldsymbol{\beta}$ is a vector of unknown parameter to estimate, τ_j is the unobservable

random cluster component, and ϵ_{in} is the "observable individual effect". The assumptions are $\tau_j \sim N(0, \sigma_\tau^2)$, and $E(\epsilon) \sim N(0, \sigma_\epsilon^2)$. From the assumption: $E(S_{in}) = \boldsymbol{X}'_{ij}\boldsymbol{\beta}$, $Var(S_{ij}) = \sigma_\tau^2 + \sigma_\epsilon^2$, and

$$\rho = corr(S_{in}, S_{in'}) = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2} = \frac{\theta^2}{1 + \theta^2}$$

where θ represents the proportion of variance explained by the cluster effect, and $\theta^2 = \sigma_\tau^2/\sigma_\epsilon^2$. The model was estimated by means of an iterative maximum likelihood procedure based on numerical integration and the EM algorithm (Dempster et al., 1977).

The multilevel ordinal logit model of Fielding et al. (2003) used the iterative generalized least squares procedures of MLwiN (Rasbash et al., 2000), and penalized quasi-likelihood in the MLwiN macros MULTICAT (Yang and Wang, 1998), incorporating the improved approximation procedures of Goldstein and Rasbash (1996).

In summary, similar to binary mutilevel probit and logit models, developing an accurate approximate method for integrating parameter and covariance components has motivated many authors' works. The most frequently used approximation methods for evaluating the integral over the random effects are Marginal Quasi-Likelihood (MQL) or Penalized Quasi-Likelihood (PQL). Both of them are available in the MLwiN software. But Rodriguez and Goldman (1995), Raudenbush et al. (2000), and Lin and Breslow (1996) have reported downwardly biased estimates using these procedures especially when the first-order Taylor expansion was used. Numerical integration that used for multilevel linear model and multilevel probit and logit models can also be used here when there are multiple categories of responses, if the assumed distribution is normal. Gauss-Hermite quadrature approximates the integral by a summation on a specified number of quadrature points Q for each dimension of the integration. If there is only one random effect, the quadrature solution requires only one additional summation over Q points relative to the fixed effect estimate. However, when there are more than five random effects, the quadrature becomes computationally burdensome.

Rabe-Hesketh et al. (2002) proposed an adaptive quadrature that works better than

MQL and PQL in multilevel latent variable models in terms of both parameter recovery and computational efficiency, since it requires fewer quadrature points to achieve the same precision. Rabe-Hesketh et al. (2005) showed through simulation that the adaptive quadrature performed well in a wide variety of situations and outperforms ordinary quadrature. The ordinary and adaptive Gauss-Hermite quadrature have been implemented in software packages including *Egret* (Corcoran et al., 1999), *Gllamm* (Rabe-Hesketh et al., 2004), $LIMDEP$ (Greene, 2002), $MIXOR$ (Hedeker and Gibbons, 1996), $MIXNO$ (Hedeker and Gibbons, 1996), *Stata* (Corporation, 2001), and SAS $PROC\ NLMIXED$ (institute, 2004).

"Maximum simulated likelihood" or "simulated maximum likelihood" method is commonly used in econometrics and transportation using simulation approach to integrate over the random effects, and is considerably faster than quatrature methods, especially as the number of random effects increase (Haan and Uhlendorff, 2006). This estimation approach has been included in LIMDEP Greene (2002) for several types of outcome variables, including nominal and ordinal variables.

Generalized linear mixed models have become increasingly popular. Many methods have been proposed for estimating such models. However, to date there is no single method that can be assumed to work well in all circumstances in terms of both parameter recovery and computational efficiency.

Bayesian Estimation to Ordered Probit/Logit Models

Several Markov Chain Monte Carlo (MCMC) algorithms (see Section 3.2.2) have been proposed for a Bayesian version of ordered probit models, among them the most notable ones are those of Albert and Chib (1993), Cowles (1996), and Nandram and Chen (1996). To avoid the bias of MLE in small samples and the inaccuracy of normal approximation, Albert and Chib (1993) first proposed a fully Bayesian approach to the binary and ordered probit model. They adopted the latent variable model in Equation 3.8, with normally dis-

tributed error terms. Then the latent variable S_i given $\boldsymbol{\beta}$ and Y_i is independent normally distributed $N(\boldsymbol{X}_i'\boldsymbol{\beta}, 1)$, and truncated to $(\gamma_{c-1}, \gamma_c]$. The observed categorical response $Y_i = c$ if $\gamma_{c-1} < S_i \leq \gamma_c$. The joint distribution of the latent variable given other parameters is formulated as:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{S}|\boldsymbol{Y}) = C \prod_{i=1}^{N} \left[\sqrt{1/2\pi} exp(-(S_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2/2) \times \left(\sum_{c=1}^{C} I(Y_i = c) I(\gamma_{c-1} < S_i < \gamma_c) \right) \right]$$

If a diffuse or flat prior is placed on $\boldsymbol{\beta}$, then $\boldsymbol{\beta}|\boldsymbol{S}, \boldsymbol{Y}$ is distributed $N\left(\hat{\boldsymbol{\beta}}, (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\right)$, where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{S}$. Albert and Chib (1993) used Markov chain Monte Carlo (MCMC) to sample from the posterior distribution for parameter estimation. Gibbs sampling allows one to sample from a multivariate posterior distribution using a full conditional distribution. For more details about Gibbs sampling, see Section 3.2.2.

Sorensen et al. (1995) presented a Bayesian analysis on genetic data of dog hip dysplasia with multiple threshold models on binary and ordered categorical data, and the models are essentially the latent variable binary and ordered probit models. They described different posterior distributions and explored the slow mixing problem of Gibbs sampling on their data. Cowles (1996) proposed a multivariate Hastings-within-Gibbs algorithm to speed up convergence when the intervals within which γ_cs must be generated from the full conditional distribution are very narrow, so there are too many bins, and this is achieved by using a Hastings algorithm (Hastings, 1970) to generate latent variables and bin boundary parameters jointly, instead of individually from their respective full conditionals.

3.1.4 Partially Ordered Models

In this section we review several partially ordered models in the literature.

Rosenbaum (1991) gave a generalization of ordered statistical methods, specifically, Wilcoxin's rank sum statistic, Spearman's rank correlation, and Page's statistic, to partially

ordered data. The method used a function of two matrices $h(\boldsymbol{X_1}, \boldsymbol{X_1})$ to measure the poset (partially ordered sets) structure of the categories and illustrated the approaches using the data of Morton et al. (1982). The study in Morton et al. (1982) focused on the lead levels in the blood of children of employees in a factory in Oklahoma who used lead in the manufacture of batteries. Employees are measured on both exposure to lead (low, medium, high) and personal hygiene (good, fair, poor). The 9 pairs of (exposure, hygiene) were seen as partially ordered. The latent lead quantity brought to home and went to children's bloodstream (x_i) was assumed, and it is decided by parents' exposure and hygiene. The response, y_i, is the lead level in children 's blood.

Meulders et al. (2005) extended the approaches by Rosenbaum (1991) for item response theory models. Item response theory (IRT) models are commonly used in education testing. In item response theory terminology, questions on a "test" are called *item*s, and the individuals taking a test are called *examinee*s. The objectives of a item response theory model are to assess the ability of examinees or the effectiveness of the test items in measuring examinees' skills or aptitude in certain topics. This skills, aptitude or certain ability are called the latent trait or latent variables, and are usually denoted by θ. An item response curve describes the probability that an individual answers an item correctly as a function of his or her latent trait θ. For ith individual and jth item, this probability is written as:

$$P(Y_{ij} = 1|\theta_i) = F(a_j\theta_i - b_j).$$

Where $F(\cdot)$ is often assumed to be a standard normal or logistic cdf, and Y_{ij} denotes the response of the ith individual to the jth item on the test, and Y_{ij}s can be binary or ordered polychotomous. The data structure for IRT models are described in Table 3.1.

Polytomous IRT models are used for items that have more than two score categories. For example, a test item that allows for partial credit, such as a rated essay question for which examinees can receive zero to four points, or a survey item with multiple ordered response levels (strongly disagree, disagree, agree, or strongly agree). Meulders et al. (2005) proposed a generalized loglinear latent variable item response model for multivariate partially ordered

	Item 1	Item 2	$\cdots$	Item k
Examinee 1	y_{11}	y_{12}	$\cdots$	y_{1k}
Examinee 2	y_{21}	y_{22}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Examinee n	y_{n1}	y_{n2}	$\cdots$	y_{nk}

Table 3.1: A typical data structure of IRT models

polytomous responses. They estimated the model with both the SAS $NLMIXED$ procedure and the MCMC Metropolis algorithm. An example of the partially ordered responses is when scoring responses to open-ended questions on different aspects, each aspect can be seen as an item, and the response to it is binary. The response combinations formed by the binary scores on the different skills (aspects) can be considered partially ordered, such as the responses coded as "000,100,010,110,001,101,011,111", where "100, 010" are less or worse than "110, 101, 011" and all are worse than "111". Thus, what they mean partially ordered is the latent level indicated by the combinations of the responses to each question. Another example they gave is their participants were rated the frequency (0 = seldom, 1 = sometimes or often) and the intensity (0 = usually mild, 1 = sometimes mild and sometimes strong, 2 = usually strong) with which they experienced four anger-related feelings. Thus, the data can be formally represented as four multicomponent posets (one for each feeling) which consist of one trichotomous and one binary component. The latent trait is measured by the frequency and intensity of experienced "anger", "irritation", "disgust" or "rage". Therefore, there are six categories resulted from combining all frequency and intensity levels ("00, 01, 10, 11, 20, 21"). The poset described by Meulders et al. (2005) has latent states associated with profiles that include a range of attributes, and a level of mastery is indicated for each of the attributes. The poset consists of the different profiles suggested by the classification of students' answers to a question. A profile j is said to be greater than a profile k if and only if for all attributes the associated mastery levels for j are at least as high as the corresponding levels for k. Thus, the profiles that include a set of responses are partially ordered. Meulders

et al. (2005) used Bayesian nonparametric density estimation methods for their generalized loglinear model with a focus on normal mixtures, and a Dirichlet process prior was adopted.

Zhang and Ip (2012) expanded the structure of poset to the partially ordered response that do not include several attributes, and developed a GLM for partially ordered data. They split the partially ordered categories into groups that contained fully ordered alternatives. They then fit ordered logit models to each of these groups. All results are then combined to yield an partially ordered model. See Ip et al. (2013) for a related mixed effect multivariate hidden Markov model to handle partially ordered disability states.

3.2 Computation

In this section, we review the most commonly used computing algorithms in non-Bayesian and Bayesian probit and logit models, as well as mutlilevel or ordered probit/logit models.

3.2.1 Non-Stochastic Numerical Methods

Numerical Integration: Laplace approximation

When the exact likelihood with multi-integrals is difficult to compute, the Laplace approximation is widely used. Suppose we wish to approximate an integral of the form:

$$\int exp\left[-q(x)\right] dx \tag{3.16}$$

by a Taylor series expansion of $q(\cdot)$ about x,

$$q(x) = q(\tilde{x}) + \frac{1}{2}q''((\tilde{x})(x - (\tilde{x}))^2 + \cdots .$$

Substitute the first two terms into Equation 3.17 yields

$$\int exp\left(-q(x)\right) dx \approx \sqrt{\frac{2\pi}{q''(\tilde{x})}} exp(-q(\tilde{x})). \tag{3.17}$$

Equation 3.17 can be extended to multivariate case. If $q(x)$ is a well-behaved function that attains its minimum value at $x = \tilde{x}$ with $q'(x) = 0$ and $q''(x) > 0$, where q' and q'' denote the gradient (the vector of first derivatives) and Hessian (the matrix of second derivatives) of q, then Equation 3.17 can be written as:

$$\int exp(-q(x))dx \approx c|q''(\tilde{x})|^{(-1/2)}exp(-q(\tilde{x}))$$

where c is a constant depending only on the dimension of the integral. In practice, as the Laplace approximation is carried to the higher order, the computational difficulty increases.

Penalized Quasi-Likelihood Estimation

The penalized quasi-likelihood can be defined as a nonlilnear regression model (Breslow and Clayton, 1993). In the case of binary data Y_i,

$$E(Y_i|\boldsymbol{\beta}, \boldsymbol{\tau}) = h(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau})$$

h$(\cdot)$ is the probit (normal CDF) or logit ($logit^{-1}$) link function. Let $\phi_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau}$, we have level-1 model:

$$Y_{ij} = \phi_i + \epsilon_i, \qquad E(e_i) = 0, \qquad Var(e_i) = \phi_i(1 - \phi_i). \tag{3.18}$$

This is a nonlinear model, and we would like to linearize it by the first-order Taylor series expansion. At the iteration t, we have:

$$\begin{aligned} \phi_i &\approx \phi^{(t)} + \frac{d\phi_i}{ds_i}(s_i - s_i^t) \\ \frac{d\phi_i}{ds_i} &= \phi_i(1 - \phi_i) = wi, \end{aligned} \tag{3.19}$$

Substituting the linear approximation $\phi^{(t)}$ for ϕ_i in Equation 3.18 gives:

$$Y_i = \phi(t_i) + w_i^{(t)}(s_i - s_i^{(t)}) + e_i.$$

Rearranging this equation so that all known terms are on the left side of the equation:

$$\frac{Y_i - \phi_i^{(t)}}{w_i^{(t)}} + s_i^{(t)} = s_i + \frac{e_i}{w_i^{(t)}}.$$

This equation can be written as the familiar two-level hierarchical linear model:

$$Y_i^{*(t)} = X_i^T \gamma + Z_i^T u_j + \epsilon_i, \tag{3.20}$$

$$\epsilon_i \sim N(0, w_i^{(t)-1}),$$

$$u_j \sim N(0, T).$$

Where,

$$Y_i^{*(t)} = \frac{Y_i - \phi_i^{(t)}}{w_i^{(t)}} + s_i^{(t)},$$

$$\epsilon_i = \frac{e_i}{w_i^{(t)}}.$$

The estimate

$$s_i^{(t)} = X_i^T \hat{\gamma}^{(t)} + Z_i^T u_j^{*(t)},$$

where $u_j^{*(t)}$ is the approximate posterior mode,

$$u_j^{(t)} = (Z_j^T W_j^t Z_j + T^{(t)-1})^{-1} Z_j^T W_j^{(t)} (Y_i^{*(t)} - X_j (\hat{\gamma})^{(t)}),$$

for

$$W_j^{(t)} = diag(w_i^{(t)}, \ldots, w_{n_j}^{(t)}).$$

The penalized quasi-likelihood algorithm works as follows:

1. Treat Equation 3.20 as a standard hierarchical linear model, using the methods of EM (Dempster et al., 1977) or Fisher scoring to maximize the likelihood with respect to the level-2 variance - covariance components T and the fixed effects γ, based on some initial estimates of s_i and w_i.

2. Based on the new estimates, $\gamma^{(t+1)}, T^{(t+1)}$, compute new weights $w_i = \phi^{(t+1)}(1-\phi^{(t+1)})$, and also compute new values of the linearized dependent variable $Y_i^{*(t+1)}$.

3. Go back to step 1. Iterate until the parameter estimates converge to some pre-assigned tolerance.

However, Breslow and Lin (1995) found that the first order penalized quasi-likelihod estimates of the variance component are seriously biased and inconsistent for multilevel logit models and proposed a bias correction. Rodriguez and Goldman (1995) also pointed out that penalized quasi-likelihood approximate procedures for estimating parameters of generalized linear multilevel models, in particular those with binary responses, can be seriously biased when the underlying random parameter values are large. Goldstein and Rasbash (1996) described a procedure which shows a considerable improvement in estimation and is implemented in currently available software. Jiang (2007) concluded that no matter to what order the Laplace approximation is carried out, the bias-corrected PQL estimator will never be consistent. Of course, as the Laplace approximation is carried to even higher order, the bias may reduce to a negligible level, but at the price of increasing computational difficulty. PQL also only works well when the variance components are small.

EM Algorithm

The EM algorithm (Dempster et al., 1977) is an iterative optimization strategy motivated by missing data and considering the conditional distribution of what is missing given what is observed. There are two main applications of the EM algorithm. The first is when the data indeed has missing values, due to problems with or limitations of the observation process. The second occurs when optimizing the likelihood function is analytically intractable, but when the likelihood function can be simplified by assuming the existence of values for additional but missing (or hidden) parameters. The latter application is more common in practice.

Assume that data $\boldsymbol{X}$ are observed and is generated by some distribution. We call $\boldsymbol{X}$ the incomplete data and assume that there is a complete data set $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{Y}$ is the

latent, unobserved variable. We also assume a joint density function:

$$f(\boldsymbol{Z}|\boldsymbol{\theta}) = f(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) = f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) L(\boldsymbol{X}|\boldsymbol{\theta}).$$

With this density function, we can define a new likelihood function:

$$L(\boldsymbol{\theta}|\boldsymbol{Z}) = L(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) = L(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}),$$

where the missing information $\boldsymbol{Y}$ is unknown and controlled by a underlying distribution, thus we can think that $\boldsymbol{X}$ and $\boldsymbol{\theta}$ are constant and $\boldsymbol{Y}$ is a random variable. The original likelihood $L(\boldsymbol{\theta}|\boldsymbol{X})$ is referred to as the incomplete-data likelihood function.

The EM algorithm first finds the expected value of the complete-data log-likelihood $log\left[L(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})\right]$ with respect to the unknown data $\boldsymbol{Y}$ given the observed data $\boldsymbol{X}$ and the current parameter estimates. Thus, we define $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to be the expectation of the joint log likelihood for the complete data, conditional on the observed data $\boldsymbol{X}$, so

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\left[logL(\boldsymbol{\theta}|\boldsymbol{Y})|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right] \\ &= E\left[log f_Y(\boldsymbol{Y}|\boldsymbol{\theta})|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right] \\ &= \int \left[log f_Y(\boldsymbol{Y}|\boldsymbol{\theta})\right] f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) dz. \end{aligned}$$

As a most general form of the EM algorithm, EM is initiated from $\theta^{(0)}$ then alternate between two steps: E for expectation and M for maximization.

1. E step: Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

2. M step: Maximize the expected value we just calculated $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$. Set $\boldsymbol{\theta}^{(t+1)}$ equal to the maximizer of Q.

3. Return to the E step unless a stopping criterion has been met.

Each iteration increases the log-likelihood, and the algorithm will converge to a local maximum of the likelihood function (Wu, 1983). A modified form called Generalized EM (GEM) is to change the M-step. Instead of maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, we find some $\boldsymbol{\theta}^{(t+1)}$ such that $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) > Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. It is also guaranteed to converge. One if the most widely used application is EM algorithm is to find the MLEs.

Newton-Raphson algorithm

Newton-Raphson (NR) algorithm also uses Taylor linearization to approximate the root (x) in a function $g(x)$. This method requires $g(x)$ to be differentiable on an interval and $g'(x) = 0$ at a root. Suppose x^* is the root for $g(x^*) = 0$, for x close to x^*,

$$\begin{aligned} 0 = g(x^*) &\approx g(x) + g'(x)(x^* - x), \\ x^* - x &= \frac{-g(x)}{g'(x)}, \\ x^* &\approx x - \frac{g(x)}{g'(x)}. \end{aligned}$$

We start from the initial guess x_0, which is hopefully close to x^*, and implement an iterative operation to find x^*:

$$\begin{aligned} x_1 &= x_0 - \frac{g(x_0)}{g'(x_0)} \\ &\dots \\ x_{i+1} &= x_i - \frac{g(x_i)}{g'(x_i)} \quad i = 0,1,2,... \end{aligned}$$

Continuing until $|x_{i+1} - x_i| < \epsilon$, for some $\epsilon > 0$, and the ending point is the approximation to the root x^*.

Convergence of NR depends on the form of $g(x)$ and the choice of the starting value, that is, how close x_0 is to x^*. If $g(x)$ has two continuous derivatives and x^* is a simple root of $g(x)$, then there exists a neighborhood of x^* for which NR converges to x^* for any x_0 in that neighborhood. In addition, if $g(x)$ is convex, ($g''(c) > 0$ for all x) then NR

converges from any starting point. But many functions are not convex, so the first result is most practical. However, it does not tell you how to find the neighborhood from which NR converges regardless of the starting value. The first result suggests convergence will occur if you start close to x^*. There are situations that NR does not converge from the designated starting value. In this cases, the distance between approximations $|x_{i+1} - x_i|$ is increasing. The Newton-Raphson method can be used to approximate the multi-dimensional integral in finding the fixed and random effects in generalized linear mixed models, but it is well known that Newton-Raphson may be inefficient and extremely slow when the dimension of the solution is high.

Normal approximation and Posterior Mode

The normal approximation is widely used by the empirical Bayes approaches to estimate parameters in the general linear mixed models including the multilevel logit or probit models. The idea of empirical Bayes can be traced back to Maritz (1970) and Robbins (1964), and the major work was established by Efron and Morris (1972, 1973, 1975). Empirical Bayes method gained popularity in 1970s and 1980s. Empirical Bayes approaches use essentially maximum likelihood estimation for a subset of the parameters. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n) \overset{iid}{\sim} f(\boldsymbol{X}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector including fixed and random effects as well as variance and covariance components. Suppose the prior $\pi(\boldsymbol{\theta})$ and $f(\boldsymbol{X}|\boldsymbol{\theta})$ are positive and twice differentiable near $(\hat{\boldsymbol{\theta}})^{\pi}$ (the posterior mode). A posterior mode is also called "generalized MLE", a maximum posteriori probability (MAP) estimate, and can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data. Then, under suitable regularity conditions, the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ for a large sample size can be approximated by a normal distribution having the mean equal to the posterior mode, and covariance matrix equal to negative inverse Hessian (second derivative matrix) of the log posterior evaluated at the mode. The matrix can be notated as $[I^{\pi}(x)]^{-1}$, the

"generalized" observed Fisher information matrix for $\boldsymbol{\theta}$. Symbolically,

$$I_{ij}^{\pi}(\boldsymbol{x}) = -\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} log\left(f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta_i})\right)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{\pi}}.$$

If the prior is flat, we replace the posterior mode $\hat{\boldsymbol{\theta}}^{\pi}$ with the MLE $\hat{\boldsymbol{\theta}}$, and the generalized observed Fisher information matrix is replaced by the usual observed Fisher information matrix, $I(\theta)$, where,

$$I_{ij}(\boldsymbol{x}) = -\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} log(f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta_i}))\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

This can simply be notated as $p(\boldsymbol{\theta}|\boldsymbol{x}) \sim N(\hat{\boldsymbol{\theta}}^{\pi}, [I(\boldsymbol{x})]^{-1})$. Wong and Mason (1985) used the normal approximation and the Newton-Raphson algorithm to estimate the vector of parameters in the multilevel logistic model since the posterior distribution of the parameters is not of closed form and was very difficult to solve at the time. Their idea is to estimate the random effect parameters with maximum likelihood estimation, plug in the estimates into the likelihood, and use standard Bayesian estimation for the remaining parameters.

3.2.2 Bayesian Computation methods

Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) methods are powerful tools in modern statistics, and constitute a collection of methods for generating samples from posterior distributions. In Bayesian statistics, Markov Cain Monte Carlo methods are used widely to numerically estimate multi-dimensional integrals. The key to Markov Chain simulation is to create a Markov process whose stationary distribution is $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ (the posterior distribution), and run the simulation long enough that the distribution of the current draws is close enough to the stationary distribution. After reaching stationarity, all further iterations of the Markov Chain are approximate realizations from the target distribution (here the posterior distribution). These realizations can be used together with Markov Chain Monte Carlo estimators to estimate parameters or summaries of the posterior distribution (Glynn and Iglehart, 1989).

When the realizations from the posterior distribution are dependent samples, according to Ergodic theorems, the approximation still works well (Karlin and Taylor, 1981).

Gibbs Sampling Gibbs Sampling is an adaption of the Metropolis-Hastings algorithm (Hastings, 1970), which is a Markov Chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult to do. It was first described by Geman and Geman (1984). Gelfand and Smith (1990) reviewed Gibbs sampling and revealed its great potential in calculating Bayesian posterior densities for a variety of structured models. Gibbs sampling allows one to sample from posterior distribution using the full conditional distributions. A full conditional distribution is the conditional distribution of a parameter given all of the other parameters in the model. The stationary distribution is approximately equal to the desired posterior distribution. The data set which is obtained by Gibbs sampling converges to the joint posterior distribution of the parameters (Gelfand and Smith, 1990). For the parameters $\boldsymbol{\theta}_i$ $(i = 1, \ldots, k)$, we set the initial values for the parameter as $\theta_1^0, \ldots, \theta_p^0$. Good starting values can be MLEs, or estimated posterior means. At the tth iteration, each element of θ_i^t is updated by sampling form the full conditional distributions:

$$\begin{aligned}
\theta_1^{(t)} &\sim \pi(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_k^{(t-1)}, \boldsymbol{y}) \\
\theta_2^{(t)} &\sim \pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)}, \boldsymbol{y}) \\
\theta_3^{(t)} &\sim \pi(\theta_3|\theta_1^{(t)}, \theta_2^{(t)}, \theta_4^{(t-1)}, \ldots \theta_k^{(t-1)}, \boldsymbol{y}) \\
\vdots &\sim \qquad\qquad \vdots \\
\theta_k^{(t)} &\sim \pi(\theta_k|\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{(k-1)}^{(t)}, \boldsymbol{y})
\end{aligned}$$

If we run a large number of iterations, the joint distribution of $\theta_i^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_k^{(t)})$ approaches the posterior distribution. We can run the algorithm as long as needed to reach the level of precision we want. The full conditionals are easy to sample from if the model is conjugate.

To reduce the influence of bad starting values, we remove the first values, and this is called burn in. Often, there is autocorrelation among the draws in each parameter sequence, while high correlation among draws in each parameter sequence makes the convergence of the sequence slow, some authors (Plummer et al., 2005; Zuur et al., 2002) recommended using thinning (only keep every kth observation for some k) to reduce autocorrelation. We take care of these issues by monitoring graphical plots of the sampled parameter values (trace plots). Gelman and Rubin (1992) described a more formal diagnostic for convergence. How many iterations do the Gibbs sampler need to run? Raftery and Lewis (1992) concluded that the required iterations for Gibbs sampler can be quite different, even for different quantities of interest within the same problem. Raftery and Lewis (1992) designed a method to test for the needed number of iterations. The examples they gave need iterations from about 5000 to 30,000 iterations. Things to note about Gibbs sampler are: first, parameters need not be updated individually, we can do them in blocks to reduce correlation and increase the convergence rate (Roberts and Sahu, 1997). For example, update all location parameters together when we know their multivariate distribution. Blocking the parameters can solve the problem that Gibbs sampling does not converge when the posterior distribution of the parameters is multi-modal. Second, the full conditional distributions need to be known or at least be amend to be reliably simulated from. If we are not able to sample from the full conditional distributions, then Gibbs sampling can not be used.

Many modifications of the Gibbs sampler have been proposed to increase the convergence rate. A Metropolis-Hastings algorithm can be used within a Gibbs sampler to speed up convergence, for example, Cowles (1996) used a multivariate Hastings-within-Gibbs algorithm when the intervals for the latent variable are very narrow. A hybrid Gibbs sampling, the Metropolis-Hastings steps within a Gibbs algorithm, can be very useful when the conditional distribution for one or more elements is not available in closed form (Givens and Hoeting, 2012).

Metropolis-Hastings Algorithm The Metropolis-Hastings algorithm (Metropolis and Ulam, 1949; Hastings, 1970) can be used when the Gibbs sampler fails, either because of poor convergence or the inability to sample from the full conditional distributions. Like the Gibbs sampler, an initial value for $\boldsymbol{\theta}^0$ must be set and this will affect the rate of convergence. At each iteration t, a proposed vector $\boldsymbol{\theta}^*$ is drawn from a symmetric proposal distribution $h(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$, symmetric here means that $h(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = h(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)$, later it was generalized to the non-symmetric case. This draw can be dependent on the value at the previous iteration $\boldsymbol{\theta}^{(t-1)}$. The probability of $\boldsymbol{\theta}^*$ from the proposed distribution $h(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ is:

$$r = min\left[\frac{f(y|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)h(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)}{f(y|\boldsymbol{\theta}^{(t-1)})\pi(\boldsymbol{\theta}^{(t-1)})h(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(t-1)})}, 1\right]$$

then we set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$ with probability $min(r, 1)$, or $\boldsymbol{\theta}^{(t-1)}$ otherwise. The algorithm is repeated until the chain reaches its stationary distribution, at least approximately, and continued until the desired precision is reached.

Good proposal distributions can greatly enhance the performance of the Metropolis-Hastings algorithm. A well-chosen proposal distribution produces candidate values that cover the support of the stationary distribution in a reasonable number of iterations, and produces candidate values that are not accepted or rejected often (Chib and Greenberg, 1995). If the full condition distributions are known for certain parameters, they may be updated utilizing a Gibbs step (Robert and Casella, 2004).

Non-iterative Monte Carlo Method

Importance Sampling Importance sampling (Glynn and Iglehart, 1989) can reduce variance and increase the efficiency of Monte Carlo algorithms for estimating integrals (Ferrenberg and Swendsen, 1988; Geweke, 1989). Suppose we want to generate simulations from the distribution $E_f(g(\theta))$, which does not have a closed form so is hard to do, we can simulate from importance density $h(\theta)$. Instead of drawing from a distribution which is complicated, importance sampling samples from a distribution that is similar to the target distribution

using the ratio of the target distribution to the similar distribution, and this ratio is called importance weight. The importance sampling estimator for expected value of $g(\theta)$ is:

$$\begin{aligned} E_f(g(\theta)) &= \int g(\theta) f(\theta) d\theta, \\ &= \int g(\theta) \frac{f(\theta)}{h(\theta)} h(\theta) d\theta, \\ &= \int g(\theta) w(\theta) h(\theta) d\theta, \\ &= E_h(g(\theta) w(\theta)). \end{aligned}$$

With Monte Carlo Estimation, $E_f(g(\theta))$ can be approximated as:

$$\begin{aligned} E_f[g(\theta)] &\approx \frac{1}{T} \sum_{t=1}^{T} g(\theta^{(t)}) \times \frac{f(\theta)}{h(\theta)}, \\ &\approx \frac{1}{\sum_{t=1}^{T} w_t} \sum_{t=1}^{T} g(\theta^{(t)}) w_t, \end{aligned}$$

Where $\theta^{(1)}, \ldots, \theta^{(-1)}$ are draws from $h(\theta)$ with Importance Weights:

$$w_t = \frac{f(\theta)}{h(\theta)}.$$

$f(\theta)$ is the posterior distribution in Bayesian statistics. The idea behind importance sampling is that instead of sampling from the posterior itself, which we do not know, we simulate from a different distribution, which we know and is close to the actual posterior, and "correct" it to account for the fact that it is not the actual posterior. The correction is done by weighting the sampled points by the ratio of the posterior to the importance distribution, both of which we should be able to evaluate numerically. Glynn and Iglehart (1989) noted that this estimator of $E_f(g(\theta))$ is consistent and can also follow central limit theorems under mild regularity conditions. A good choice of $h(\theta)$ can ensure a low sampling variance. The tails of $h(\theta)$ should be at least as heavy as $f(\theta)$ for accurate estimation. If this requirement is not met, then some standardized importance weights will be huge, then the few draws from importance

distribution with much higher density under $f(\theta)$ than under $h(\theta)$ will receive huge weights and inflate the variance of the estimator. As it is shown above, importance sampling can be used to estimate posterior distributions that are only known up to a constant of ratio. Importance sampling estimators save time when we have many models under consideration; we can use a single sample from one of the models to give a rough approximation under each model.

Note on Possibly High Dimensions

When the number of variables gets larger in the joint distribution, it becomes difficult to calculate the marginal distribution by integration. Suppose there is a joint distribution $f(x, y_1, y_2, \ldots y_n)$, then the marginal distribution of x is:

$$f(x) = \int \cdots \int f(x, y_1, y_2, \ldots y_n) dy_1 \ldots dy_n \tag{3.21}$$

it is very difficult to integrate out all other variables in equation 3.21 when there are many of them. In a Bayesian setting, for a posterior distribution with parameter $\boldsymbol{\theta}$ (a $n \times 1$ vector),

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{L(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}. \tag{3.22}$$

When the value of n is very large, calculating the marginal distribution is not feasible, and this is the "curse of dimensionality" in finding posterior distributions. Before Gelfand and Smith (1990) brought Gibbs sampling to Bayesian statistics, Wong and Mason (1985) found that as the dimension of the parameter increases, numerical integration became more difficult to achieve accurate approximation without an astronomical amount of calculation.

Markov Chain Monte Carlo methods like Gibbs sampler enable us to avoid calculating high dimensional integrals by sampling a series of one dimensional random variables. Gibbs sampler makes sampling from the posterior distribution much easier as the dimension of the problem increases. Gelfand (2000) explained how to avoid the curse of dimensionality: suppose $P(\theta \longrightarrow A)$ is the transition kernel of a Markov chain with stationary distribution

$h(\theta)$, if $h^{(0)}(\theta)$ is a density that gives starting point for the chain, then with this staring point $\theta^{(0)}$ and using $P(\theta \longrightarrow A)$, we can generate a trajectory of the chain: $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}$. If t is large enough, $\theta^{(t)}$ is approximately from the target distribution $h(\theta)$. Raftery and Lewis (1992) gave some examples to sample from a 10-dimensional multivariate normal posterior distribution and an 190-dimensional posterior distribution from spatial statistics using Gibbs sampler. They investigated how many iterations we have to run Gibbs sampler to make it converge to the target distribution, and reach the desired level of precision.

Sampling from truncated distributions

Gelfand et al. (1992) described using Gibbs sampling to sample from truncated distributions to carry out a full Bayesian calculation. These kinds of complex constrained parameter problems were previously unanalyzable by standard numerical integration techniques. Examples were given about sampling from ordered multinomial exponential family parameters and other data with truncated restrictions. The introduction of a latent variable reduces sampling from the truncated densities such as truncated normal, beta, and gamma, to the sampling of two or several uniform random variables, and is easier to code. Polson (1993) found that including a latent variable in log-concave densities can improve convergence.

Damien and Walker (2001) discussed in detail about a slice sampling for a truncated normal distribution. Use a simple example, suppose $X \sim N(0,1)$, that is:

$$f_X(x) \propto exp(-\frac{x^2}{2}).$$

Introduce the latent variable S which has joint distribution with X given by

$$f_{X,S}(x,s) \propto I_{(0,exp(-\frac{x^2}{2}))}(s),$$

then we have the full conditional distribution:

$$\begin{aligned} S|X=x &\sim U(0, exp(-\frac{x^2}{2})), \\ X|S=s &\sim U(-\sqrt{-2log(s)}, \sqrt{-2log(s)}). \end{aligned}$$

It is simple to implement this idea for a truncated standard normal variable. Suppose we wish to sample from the truncated standard normal density given by

$$f_X(x) \propto exp(-\frac{x^2}{2}) I_{(x\in(a,b))}.$$

The joint distribution of latent variable S and X is:

$$f_{X,S}(x,s) \propto I_{(0,exp(-\frac{x^2}{2}))}(s) I_{(x\in(a,b))}.$$

which leads to the full conditionals:

$$\begin{aligned} S|X=x &\sim U(0, exp(-\frac{x^2}{2}), \\ X|S=s &\sim U(max(a, -\sqrt{-2log(s)}), min(b, \sqrt{-2log(s)})). \end{aligned}$$

This algorithm added one more full conditional: a uniform distribution to the Gibbs sampling. Damien and Walker (2001) introduced a novel approach to sampling truncated densities: Adaptive Uniform Rejection Sampling (AURS), which is similar in spirit to Adaptive Rejection Sampling by Gilks and Wild (1992). The AURS algorithm is: let $h(\cdot)$ be a non-increasing continuous function on $(0, c)$, such that $0 < h(0) < \infty$. To sample from $f(u) \propto h(u)$, at the ith iteration of the algorithm we have evaluated $h(u_j)$ at u_j for $j = 1, \ldots, i+1$, where $0 = u_1 < \cdots < u_i < u_{i+1} = c$. Let

$$g_i(u) \propto \sum_{j=1}^{i} h(u_j) I(u_j < u < u_{j+1})$$

Take u' from $g_i(\cdot)$ and w from the uniform distribution on the interval (0, 1). If $w < \frac{h(u')}{h(u_j)}$ where $u_j \leq u' < u_{j+1}$, then we accept u' as a random variate from $f(u)$. Otherwise, we proceed to the $(i+1)$th iteration with $g_{i+1}(\cdot)$. For more information and algorithms about sampling from truncated densities, see Robert (1995) and Foulley (2000).

Chapter 4

Bayesian Partially Ordered Probit and Logit Models with an Application to the Course Redesign in Fall 2012

4.1 Problem Statement

This chapter presents Bayesian partially ordered probit and logit models and used these models to evaluate the impact of Intermediate Algebra course redesign at the University.

Multilevel models (i.e., random effect or hierarchical models) have become a popular choice in education research. Raudenbush (1988) reviewed statistical methods in educational statistics and concluded that multilevel linear models were a prominent theme in modeling education data. Kaplan and Elliott (1997) summarized that the application of multilevel linear regression methods had resulted in extraordinary advances in school process research since multilevel linear models account for the clustered sampling schemes in education research. Their popularity has increased over time. Similar to simple linear regression, the response data for multilevel linear models should be continuous and the error

terms are assumed to be normally distributed. However, the response data in education (student success) is often recorded as students' letter grade A, B, C, D (possibly with +/-) and F, as well as CR/NC (Pass/Fail), W (Withdrawal), I (Incomplete) and AUD (Audit). Obviously, it is not a case to fit multilevel linear models.

Meanwhile, common models used for categorical response data (such as students' letter grades) are probit/logit models (Aldrich and Nelson, 1984; Wong and Mason, 1985; Stiratelli et al., 1984) and ordered probit/logit models (Schaafsma and Osoba, 1994; Fullerton, 2009; Hedeker, 2008). These approaches enable a researcher to model binary or ordinally scaled dependent variables with one or more independent variables. The response data for ordered probit/logit models has to be ordered with about equal distance between neighboring points. All of these models can be easily fit in software packages such as SAS ($PROCGLIMMIX, NLMIXED, LOGIT$ and others) and R (*nlme* and *polr* libraries and others). Maximum likelihood Estimation is primarily used by these procedures.

However, we often have data that does not fit into one of these scenarios. For example, assume that the exam score or course score of students were not recorded, and we observe students' letter grade with possible values: A, B, C, D, F, CR (Pass with Credit), NC (Fail with No Credit), I (Incomplete), AUD (Audit the course) and W (Withdrawal). This variable is partially ordered because, while an A is clearly better than a D, a CR is not necessarily better or worse than an A, B, or C. A multilevel/hierarchical linear regression or a simple linear regression model clearly do not fit this categorical response data. In order to fit a multilevel binary probit/logit model to this type of data, we have to collapse all passing grades together and all failing grades together. Moreover, it is hard to classify I, AUD, W to either pass or fail. Similarly, to fit an ordered probit/logit model, we have to only consider the fully ordered letter grade (A-F), but give up the unordered data. Both of these approaches result in a loss of information by either truncation or exclusion. An unordered probit/logit model ignores the ordered information among the letter grades, and does not answer our research question, see Section 4.4 for more explanation.

For computational reasons, we consider Bayesian models in this chapter. Albert and Chib (1993) first brought a Bayesian method using a latent variable to model dichotomous and polychotomous ordered and unordered response data. In this dissertation, we expand the model of Albert and Chib (1993) to partially ordered response data with random effects in a fully Bayesian approach. We also extend the results to a logit model using the idea of O'Brien and Dunson (2004). We apply these models to an example of the course redesign evaluation at the University.

4.2 Data and Descriptive Statistics

4.2.1 Data and variable introduction

The initial data file consisted of available demographic information and all courses taken by the 1308 students who were enrolled in Intermediate Algebra in Fall 2012. Each student enrolled in one or several courses, and some registered in multiple sections of some courses. The data has the section number of each course, number of credit hours and a letter grade for each course, college, college description, department, race/ethnicity, gender, high school they graduated from, high school graduation date, ACT score, SAT score, and several other variables. For duplicate courses, i.e., when a student registered for the same course in two or more sections and received two or more grades, we kept the highest grade for the same course.

We were limited to the student data approved for use by the Institutional Review Board. Since we had all the courses each student enrolled in the Fall semester of 2012, we used this information to calculate students' course load (total credit hours each student took except Intermediate Algebra) in Fall 2012, and the semester GPA of each student in the semester when they took Intermediate Algebra, excluding Intermediate Algebra itself. It is common sense that if a student takes many courses in the semester, he or she will not have

as much time to study for each course, and the grade in Intermediate Algebra would likely be correspondingly lower. On the other hand, a student with a lower course load should tend to do better in each course. We expect that a students' course load will be a good predictor variable for students' Intermediate Algebra success.

Likewise, the semester GPA will reflect all of the outside factors that might affect a student's performance in Intermediate Algebra. For example, if a student is working many hours at a job or has a serious illness, that is likely to affect his or her grades in all courses, not just Intermediate Algebra. The semester GPA will also reflect non-cognitive skills like study habits, tendency to attend or skip classes, and so on, that can impact academic success across the board. So we included the semester GPA as a proxy measure for the combination of factors outside of the teaching environment that are likely to impact success in Intermediate Algebra.

SAT and ACT scores are a strong predictor of success in lower division courses in college. They have been used by most of the universities as the most important criterion for college admission and are used for placement into lower-division mathematics courses at the University; students with a math SAT score between 450 and 500 or math ACT score between 19 and 21 are placed into Intermediate Algebra. Since students' math SAT or ACT score reflect some combination of mathematical aptitude and achievement, we included this variable in the analysis.

Many students take either the SAT or the ACT but not both. Because both SAT and ACT math scores measure a similar construct, we combined them into a single variable SAT/ACT using the national mean and standard deviation provided by the companies who produce those assessments to calculate standardized z-scores. That is, using the ACT Math or SAT math score of each student minus the mean ACT or SAT score of all students in the U.S., and divided by the standard deviation of ACT or SAT score of all students in the U.S., denoted SAT/ACT in this work. If a student took both of the exams or the same exam more than once, the highest score is assigned to SAT/ACT.

Most students in college have taken and passed a course in high school that addresses the content in Intermediate Algebra. A problem is that it may have been many years since they learned the material. We made two variables to describe this behavior. "High school" is an indicator for the students who have a high school diploma or were still in high school when they were enrolled in Intermediate Algebra at the University. Students without a high school diploma, but have a GED certificate may include home schooled students, or any other non-high school graduates who found their way into the University. The time in years since high school graduation (named $HS\ grad.\ years$) is calculated as the elapsed time in years between the day students graduated from high school to August 20, 2012, when the Fall semester of 2012 began. These two variables allow us to include information on the students' scholarly past and how recent it is.

Gender and race/ethnicity (self-identified) were used as control variables in our work. We considered the groups: Asian, white, Hispanic, native American and "other races". "Other races" are used as reference group in the models.

There are 11 different instructors who taught the traditional lecture sections in Fall 2012. The 12th instructor was in charge of all the students in the Redesigned sections, and did not teach any of the traditional lecture sections of the course in the Fall semester of 2012. This is the pivotal independent variable in our work. The instructor variable was treated as a random effect in all models used in this work. Thus, we can compare the Redesigned sections with the traditional lecture sections by making inferences about the instructor random effects.

In the process, we found that the variable about college and department:
$COLLEGE\ DESC$ and $DEPARTMENT\ DESC$ are actually which department and college offered the course, not the college or department a student belongs to. For example, Intermediate Algebra are always offered by Math and Statistics department and College of Art and Science. Thus, the information of $COLLEGE\ DESC$ and $DEPARTMENT\ DESC$ is useless, and tells us nothing about students. We removed the information for other courses after we calculated $Semester\ GPA$ and $Course\ load$ because we are interested in students'

Variable Name	Description
SAT/ACT	Standardized SAT or ACT math score.
Course load	Total credit hours each student took in the Fall semester of 2012, excluding Intermediate Algebra.
Semester GPA	A weighted average of students' GPA in the Fall semester of 2012, calculated excluding Intermediate Algebra.
High school	"1" if a student graduated from high school, "0" if a student has a GED certificate.
HS grad. years	The elapsed time in years between a student's high school graduation date and Aug. 20, 2012.
Gender	Male/Female.
Race /Ethnicity	Self identified race/ethnicity of a student. Restricted to Asian, White, Hispanic, Native American or of other race/Ethnicity.
Instructor	Which instructor taught the section the student was enrolled in. Instructors are labeled with numbers 1-12, 12 denotes the Redesigned course.

Table 4.1: Detailed descriptions of independent variables.

Intermediate Algebra success in this dissertation, not what courses a department offers.

The predictor variables used in this dissertation are described in Table 4.1.

4.2.2 Descriptive Statistics

In this sections, we examine the impact of the independent variables on student success in Intermediate Algebra through a series of plots, tables and tests.

Distribution of independent variables

We made plots to learn about the distribution of the quantitative independent variables: *HS grad. years*, SAT/ACT, *Course Load* and *Semester GPA*.

The upper left corner of Figure 4.1 displays a histogram of $HS\ grad.\ years$. We have defined $HS\ grad.\ years$ as the elapsed time in years between students graduated from high school and August 20, 2012, the first day of the Fall semester when they took Intermediate Algebra. Some students are still in high school making their $HS\ grad.\ years$ negative. We see that about 70% students just graduated from high school when they took Intermediate Algebra in the Fall semester of 2012, while only a few had graduated from high school for 15 or more years. The mean value of $HS\ grad.\ years$ of students in Fall 2012 is 2.602 years.

The upper right corner Figure 4.1 describes a histogram of SAT/ACT (standardized SAT or ACT math score). The histogram shows that SAT/ACT is roughly normally distributed, the mean standardized SAT/ACT score is -0.458 (ACT = 19, SAT = 463), and the standard deviation is 0.4239. For all U.S. ACT and SAT math scores, mean of ACT = 21, Std. Dev. of ACT = 5.3; mean of SAT = 516, Std. Dev = 116. Recall that the placement score to Intermediate Algebra is: ACT math 19-21, SAT math 450-500.

The lower left corner Figure 4.1 provides a histogram for the variable $Course\ load$, which is the credit hours students took excluding Intermediate Algebra in Fall 2012. About 20% of the Intermediate Algebra students have a $Course\ load$ 12 credit hours, and around 45% of students had 15-16, 10% of students had 18 or more. According to the student statistics in the Fall semester of 2012 reported by the registrar office of the University, the average load of undergraduate students is 12.86. The average course load of students enrolled in Intermediate Algebra course is actually 16.97, which is 13.97 plus the three credit hours in Intermediate Algebra.

The lower right corner of Figure 4.1 gives a histogram for the variable $Semester\ GPA$. The histogram tells us that the semester GPA of Intermediate Algebra students in the Fall semester of 2012 is approximately normally distributed with mean 2.97 and standard deviation 0.656.

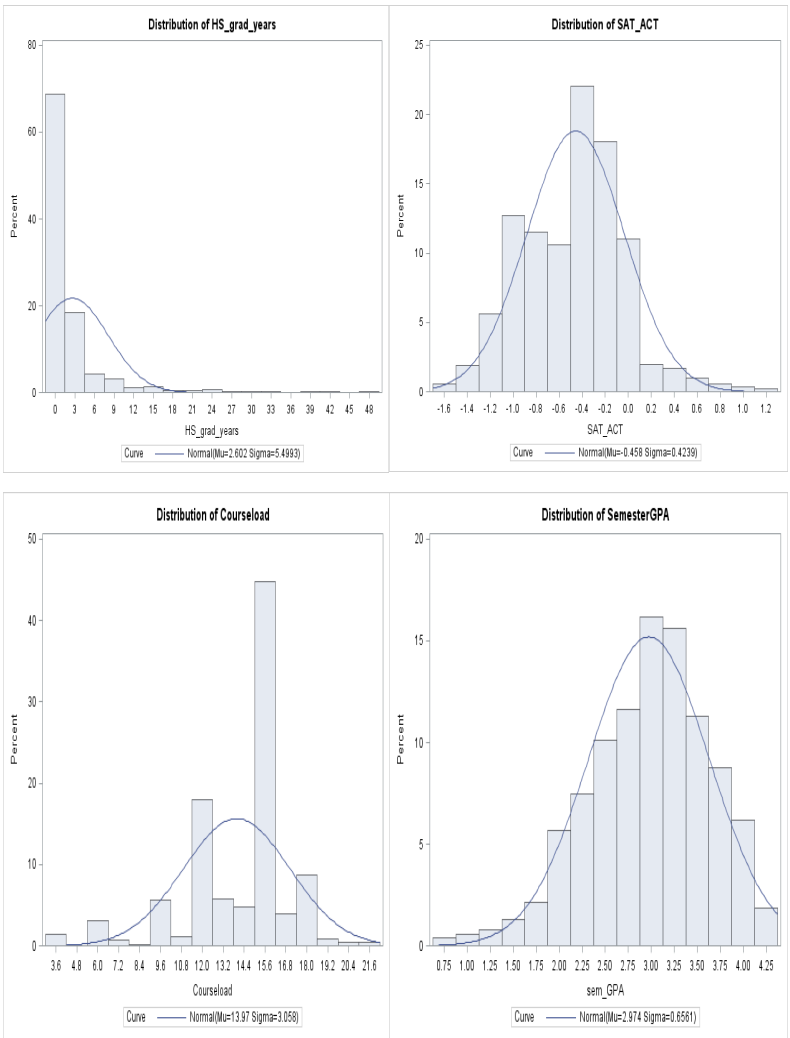

Figure 4.1: Histograms for quantitative independent variables about Intermediate Algebra students in the Fall semester of 2012. The left upper panel displays the distribution of the elapsed time in years after students graduated from high school to the first day of the Fall semester of 2012; the right upper panel: the distribution of SAT/ACT score of the students enrolled in Fall 2012. Left lower panel presents the distribution of the total credit hours each student took. The right lower panel gives the distribution of students' semester GPA in the Fall semester of 2012.

Student success between the two teaching methods: the Traditional lecture and the Redesign

First, we would like to roughly examine students success measured by pass rate between the two teaching methods: the Traditional lecture and the Redesign with a standard frequency table. Table 4.2 describes student success between the two teaching methods. We have to decide to treat "I" (Incomplete), "W" (withdrawal), "AUD" (Audit) as a pass or a fail. If we treat the "I" (Incomplete), "W" (withdrawal), "AUD" (Audit) as a fail, the pass rate for traditional lecture sections is 62.09% (678/1092), and the pass rate for the Redesign students is only 30.56% (66/216).

The Chi-Square test is a commonly used method to test for association between categorical variables. It was created by Pearson (1900) and later refined by Fisher (1948). We use the Chi-Square test to investigate if there is relationship between students' pass rate in Intermediate Algebra and the teaching method. Alternative, we can use Fisher's exact test. The Chi-Square test can safely be used with critical values from a Chi-square distribution when no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater. In particular, all four expected counts in a 2×2 table should be 5 or greater. The p-value is the area to the right of the test statistic under the density curve of the Chi-square distribution with $(r-1)(c-1)$ degrees of freedom, and r represent the number of rows, c is the number of columns of the two-way table. In this case, the null and alternative hypothesis that we are testing by the Chi-Square test are:
H_0: The pass rate of students in the traditional lecture sections = the pass rate of students in the Redesign
H_1: The pass rate of students in the traditional lecture sections $\neq$ the pass rate of students in the Redesign

The Chi-square test statistic with 1 degree of freedom is 73.10 in our case, with a p-value

<0.0001. This suggests that the pass rate of Intermediate Algebra students between the traditional lecture sections and the Redesign is statistically different.

	Traditional Lecture	Redesign	Total
Passed students	678	66	744
Passed percentage among all students	51.83	5.05	56.88%
Pass rate in each group	**62.0**	**30.56**	
Failed students	414	150	564
Failed percentage among all students	31.65	11.47	43.12%
Failing rate	37.91	69.44	
Total number of students	1092	216	1308
Percentage of students in each group over all	83.49	16.51	100%

Table 4.2: A comparison of the pass rate between traditional lecture sections and the Redesign. The "Passed percentage among all students" means the percentage of passed students by each teaching method over all students enrolled in the Intermediate Algebra course in the Fall semester of 2012, similarly, the "Failed percentage among all students" means the percentage of failed students taught by each teaching method over all students enrolled in the Intermediate Algebra course in the Fall semester of 2012. The "Pass (Failing) rate" represents the percentage of passed (failed) students from traditional lecture or the Redesign when I, W, AUD are treated as a fail.

Grade distribution of students in the Redesign

Because the pass rate of students in the Redesign is very low, we would like to look into the letter grades of the Redesign students. Table 4.3 describes the distribution of students' grades in the Redesign. 54.63% (118/216) of students got an I (incomplete) grade. We treated the "I"s as failing the course. That is why the non-passing rate in the Redesign is so high. However, these 54.63% of students who were given Incomplete are still continuing in the Spring semester, and students in the Redesign are allowed to have three semesters

(including summer) to finish the course, and get the "I" changed to a passing grade. The "I" will become an "F" if a student still can not passed the course at the end of the third semester.

Letter Grade	Frequency	Percent	Cumulative Frequency
I	118	54.63	118
W	30	13.89	148
A+	14	6.48	162
A	26	12.04	188
A-	14	6.48	202
B+	6	2.78	208
B	4	2.41	212
B-	1	0.46	213
C	1	0.46	214
D	2	0.92	216

Table 4.3: Grade distribution of students in the Redesign in Fall 2012

Student success in two populations: students with an SAT/ACT score and those without an SAT/ACT score.

Either a SAT or an ACT score are required by many universities as the one of the criteria for college admission, and is thought widely as related to students' learning ability and academic performance in college. In the data, there is a small portion of students who did not take the SAT or ACT and still enrolled in Intermediate Algebra. We would like to see how this group of students perform in Intermediate Algebra relative to those who took the SAT or ACT test. Table 4.4 provides a comparison of students success between two groups. The first row gives the number of students who passed Intermediate Algebra from the two groups: there were 700 students who took the SAT or ACT passed the course, and only 44 students not taking a SAT or an ACT passed the course. The second row starting with "Pass rate among all students" gives that the overall pass rate in Intermediate Algebra in

the Fall semester of 2012, which is 56.88%. Of the 56.88% of students who passed, 53.52% of them took a SAT or ACT test, and 3.36% did not. The third row "Pass rate in each group" gives the percentage of students who passed the course in each group (those with the SAT/ACT and those without): 57.85% of students who took an SAT/ACT passed the course, while percentage of passed students among those who did not take the SAT/ACT was lower (44.90%). Similarly, the next rows provide information about students who did not pass the course from the two groups: either took the a SAT or an ACT or did not. The last two rows under "Total number of students in each group" gives the total number of students who took the SAT or ACT (1210 students) and those who did not take (98), as well as the percentage: 92.50% of students enrolled in the Intermediate Algebra in Fall 2012 took the SAT or ACT, only 7.49% students in the course in that semester did not take the SAT or ACT test, and there were total 1308 students enrolled for Intermediate Algebra after the first three weeks of the semester.

According to the University's admission policy, students are required to have an SAT or ACT score to be admitted to the University, unless they have at least 24 college credits. These 98 students who did not have a SAT or an ACT score are likely transferred students from other universities. For example, some private universities do not require students to have a SAT or an ACT test score. Alternatively, these students might have taken the compass placement test.

The Chi-Square test statistic is 6.2019, the corresponding p-value is 0.0128, which is smaller than the standard 0.05 significance level. Thus, we conclude that the pass rates of students who have an SAT/ACT score is statistically different from those without an SAT/ACT score. We will further examine how the students with an SAT/ACT score and those without are assigned into the traditional lecture sections and the Redesign.

	Students who took SAT/ACT	Students who did not take SAT/ACT	Total
Passed students	700	44	744
Passed percentage among all students	53.52	3.36	56.88%
Pass rate in each group	**57.85**	**44.90**	
Failed students	510	54	564
Failed percentage among all students	38.99	4.13	43.12%
Failing rate in each group	42.15	55.10	
Total number of students in each group	1210	98	1308
Percentage of students in each group over all	92.51	7.49	100%

Table 4.4: the pass rate comparision between students took an SAT/ACT and those who did not in Fall 2012. The "Passed percentage among all students" means the percentage of passed students in each group over all students enrolled in the Intermediate Algebra course in the Fall semester of 2012, similarly, the "Failed percentage among all students" means the percentage of failed students in each group over all students enrolled in the Intermediate Algebra course in the Fall semester of 2012. The "Pass (Failing) rate" represents the percentage of passed (failed) students in each group (students with an SAT/ACT score or those without enrolled in the course in Fall 2012) over all students in each group.

How students with and without an SAT/ACT score were split up into the Traditional lecture sections and the Redesign?

The results in Section 4.2.2 suggest that students with an SAT/ACT did significantly better than those without an SAT/ACT score. Since we are concerned with students success in the Redesign, we would like to know the proportion of students with an SAT/ACT and those without an SAT/ACT in the Redesign. Note that this is an observational study, and students are not randomly assigned into the Traditional lecture sections (our control group) or the Redesign (our treatment group).

Table 4.5 describes the component of students (with or without an SAT/ACT score) in the Traditional lecture sections and the Redesign. The four rows after "Traditional lecture" tell the number and proportion of students who took the SAT/ACT in the Traditional lecture section: among the 1092 students in the Traditional lecture section of Intermediate Algebra course in Fall 2012, 1018 of them took the SAT/ACT, and 74 of them did not take the SAT/ACT. 83.49% of students were in the Traditional lecture sections, among those, 77.83% took an SAT/ACT, while 5.66% did not. Among all students in the Traditional lecture sections, 93.22% of students took the SAT/ACT test, only 6.78% of students did not take SAT/ACT test. The row starting with "percentage 3" tells that 84.13% of students who took the SAT/ACT in the Intermediate Algebra class were in the Traditional lecture sections, while only 75.51% of students without the SAT/ACT score were in the Traditional lecture sections. The four rows after "Redesign" gives students' combination in the Redesign. We see that only 15.87% of students with the SAT/ACT registered the Redesign, while there were 24.49% of students without SAT/ACT score in the Redesign. The percentage of students in the Redesign among all Intermediate Algebra students is 16.5%. The row beginning with "percentage 5 in the Redesign" telling us that among the students in the Redesign, 89.89% of them have SAT/ACT score, while 11.11% of them do not. The corresponding percentages in the Traditional lecture section are 93.22% and 6.78%.

The p-value from the Chi-Square test of the data in Table 4.5 is 0.027, suggesting that the proportion of students with an SAT/ACT score registered the Redesign is significantly lower than that of those without the SAT/ACT score (15.87% vs. 24.49%) in the Traditional lecture sections. Also, the proportion of students with an SAT/ACT score is significantly higher than that of those without an SAT/ACT score (84.13% vs. 75.51%) in the Traditional lecture sections. In other words, considering the students who do not an SAT/ACT score have a significantly lower pass rate than those who do. This also means that more unprepared students went to the Redesign. Generally comparing only the pass rate between Traditional lecture sections and the Redesign creates a bias in favor of Traditional lecture sections.

	students took SAT/ACT	students did not take SAT/ACT	Total
Traditional	1018	74	1092
percentage 1 among all students	77.83	5.66	83.49%
percentage 2 in lecture sections	93.22	6.78	100%
percentage 3 in each category	**84.13**	**75.51**	
Redesign	192	24	216
percentage 4 among all students	14.68	1.83	16.51%
percentage 5 in the Redesign	88.89	11.11	100%
percentage 6 in each category	**15.87**	**24.49**	
Total	1210	98	1308
percentage 7 in each category over all	92.51	7.49	100%

Table 4.5: Percentages of students with or without an SAT/ACT score in the Redesign. "percentage 1" represents the persentage of students in each group among all Intermediate Algebra students in Fall 2012, and "pencentage 2" gives that percentage of students in each group among all students in the Traditional lecture sections. The row starting with "Percentage 4" provides the pencentage of students in each group among all students in the course in the semester. The row began with "percentage 6" displays the percentage of students from each group in the Redesign. The row of "percentage 7" gives the percentage of students from each group over all students in Intermedaite Algera course in Fall 2012.

Students' Intermediate Algebra success between different race/ethnicity groups

It is known that students from different race/ethnicity groups do not tend to do the same academically. Considering the diverse nature of the University, we would like to investigate how different racial group of students do in the Intermediate Algebra course. Students' race and ethnicity are self-identified. There were only 5 non-resident aliens and 1 native Hawaiian so these races are ignored for the sake of Chi-square test because the sample

size of these two groups are too small, and we also removed the 17 students who did not provide their race/ethnicity information ("race/ethnicity unknown"). Table 4.6 provides students' success in Intermediate Algebra across different race/ethnicity groups. The row after "Pass" tells the number of students who passed the course from different race/ethnicity groups, the row starting with "percentage 1" gives the percentage of passed students in each race/ethnicity group among all students (1285). "percentage 2" represents the percentage of passed students from each race/ethnicity group among all passed students; and the row of "pass rate" provides the pass rate of students from each race/ethnicity group in Intermediate Algebra course. Similarly, the row starting with "Failed students" gives that number of students who did not pass the course; the row beginning with "percentage 3" tells the percentage of students who did not pass the course from each race/ethnicity group over all the students. "percentage 4" represents the percentage of the students who did not pass from each race/ethnicity group among all the non-passed students. The row of "Failing rate" gives the percentage of non-passed students in each race in Intermediate Algebra course. The last two rows after "Total" gives the total number of students from each race as well as percentage of students from each race in Intermediate Algebra course in Fall 2012.

The pass rate of students was the highest for Asian students (66.67%), the second highest was whites (63.82%), then the pass rate of students from "Two or more races" (56.86%), Hispanic (55.41%), native American (44.44%), then Black or African American (42.86%). The Chi-square test statistic with degree of freedom 5 is 19.29, p-value is 0.0017, which is smaller than 0.05 significance level. We conclude that the proportion of students who passed Intermediate Algebra across different race/ethnicity groups differs. Because of the different performance of students from various race/ethnicity group in Intermediate Algebra course, we should incorporate this information into the models of comparing student success between the Traditional lecture sections and the Redesign, and should not ignore this difference but only rely on the overall pass rate to give a conclusion. Again, this is because we have a non-randomized observational study, and not an experiment.

According to the report from the University registrar "Headcount by race/ethnicity and level Fall 2012", there were 21,008 undergraduate students enrolled in the University, among them, 42.96% Hispanic students, 38.29% whites, 3.05% Asian, 6.39% native American, 2.73% African American (UNM, 2012). In Table 4.6, "percentage 5" represents the percentages of students from different race/ethnicity group in Intermediate Algebra, they are: 51.83% Hispanic, 30.97% whites, 6.30% Native American, 4.36% African American, and 2.57% Asian. That means that more than average of Hispanic, and African American students were enrolled in Intermediate Algebra in Fall 2012, lower than average of white students registered in this course, and the proportion of native American and Asian students in Intermediate Algebra course were about the same as the proportion of students of these race/ethnicity groups at the University in undergraduate programs during Fall 2012.

	Hispanic	White	Native American	Two or more races	Black or Afri. Amer.	Asian	Total
Passed students	369	254	36	29	24	22	734
percentage 1	28.72	19.77	2.80	2.26	1.87	1.71	57.12
percentage 2	50.27	34.60	4.90	3.95	3.27	3.00	
pass rate	**55.41**	**63.82**	**44.44**	**56.86**	**42.86**	**66.67**	
Failed students	297	144	45	32	22	11	551
percentage 3	23.11	11.21	3.50	1.71	2.49	0.86	42.88
percentage 4	53.90	26.13	8.17	5.81	3.99	2.00	
fail rate	44.59	36.18	55.56	43.14	57.14	33.33	
Total	666	398	81	51	56	33	1285
percentage 5	51.83	30.97	6.30	3.97	4.36	2.57	100.0

Table 4.6: The number and percentage of passed or non-passed students from different race/ethnicity groups.

Student success in Intermediate Algebra course between high school graduates and GED holders

GED certificate holders include the students who dropped out of high school, or home schooled students, but passed the "General Educational Development (GED)" tests and were certified that they have American or Canadian high school-level academic skills. Since home schools can not be controlled by education administration authorities, the quality of instructors and curriculum vary greatly among home schools. We would like to analyze how these students do in Intermediate Algebra course compared to those who graduated from a regular high school.

The population for GED holding students is small at the University: for the Intermediate Algebra classes in the Fall semester of 2012, there were only 53 students had GED certificate. Table 4.7 gives a frequency table comparing the pass rate of high school graduates and GED certificate holders in the Intermediate Algebra class in Fall 2012. We see from the table that the pass rate of high school graduates was 57.93%, while the pass rate of GED certificate students was only 32.08%. The Chi-Square test statistic is 13.858 with 1 degree of freedom, and yields a p-value of 0.0002. Based on this small p-value, we conclude that the pass rate in Intermediate Algebra class between high school graduates and GED certificate students is significantly different.

Are boys better at math than girls?

It has been a misconception for a long time that men do better in math than women. To satisfy our curiosity, we investigated this phenomenon in this section. Table 4.8 shows that the pass rate was 59.75% for female students, and 53.45% for male students in the Intermediate Algebra course in Fall 2012. The Chi-Square test statistic is 5.2522 with 1 degree of freedom, and the p-value for this test is 0.0219, which suggests that the pass rate for female students in Intermediate Algebra course was statistically significantly higher than

	High school graduates	GED certificate holders	Total
Passed students	727	17	744
Passed percentage among all in the course	97.72	2.28	100%
Pass rate in each group	**57.93**	**32.08**	
Failed students	528	36	564
Failed percentage among all in the course	93.62	6.38	100%
Failing rate among each group	42.07	67.92	
Total number of students in each group	1255	53	1308
Percentage of students in each group over all	95.95	4.05	100%

Table 4.7: Passe rate between high school graduates and GED certificate holders in the Fall semester of 2012. The "Passed percentage among all students" means the percentage of passed students in each group over all students enrolled in the Intermediate Algebra course in the Fall semester of 2012. Similarly, the "Failed percentage among all students" means the percentage of failed students in each group over all students enrolled in the Intermediate Algebra course in the Fall semester of 2012. The "Pass (Failing) rate" represents the percentage of passed (failed) students in each group (high school graduates or GED certificate holders) enrolled in the course in Fall 2012 over all students in each group.

that of males. Meanwhile, from the University registrar report, 43% of all students at the University were male, and 57% of them were female in Fall 2012. In Intermediate Algebra, there were similarly more females (54.51%) than males (45.49%).

The relationship between students' Intermediate Algebra success and the amount of time since graduating from high school

Many instructors have an impression that very old students tend to do better than young students. In most cases, students' age is approximately the amount of time after they graduated from high school plus 18. We wish to see how the amount of time since students

	Female	Male	Total
Passed students	426	318	744
Passed percentage among all students	32.57	24.31	56.88%
Pass rate of each gender	**59.75**	**53.45**	
Failed students	287	277	564
Failed percentage among all students	50.89	49.11	100%
Failing rate of each gender	40.25	46.55	
Total number of students of each gender	713	595	1308
Percentage of students of each gender over all	54.51	45.49	100.00

Table 4.8: The pass rate between female and male students in the Fall semester of 2012. The "Passed percentage among all students" means the percentage of passed students in each gender over all students enrolled in the Intermediate Algebra course in the Fall semester of 2012; similarly, the "Failed percentage among all students" means the percentage of failed students of each gender over all students enrolled in the Intermediate Algebra course in the Fall semester of 2012.

graduated from high school is related with students' Intermediate Algebra success.

We split up the years since students graduated from high school into several categories: students who are still in high school, newly high school graduates who graduated from high school less than 5 years ago, students who have graduated from high school in 5 to 10 years (many students in this group were in military for several years, and went to college after they left the military), students who graduated from high school more than 10 years to 20 years ago (many people in this group took Intermediate Algebra trying to change their career), older people who have graduated from high school for 20 years to 30 years, as well as the students who have graduated from high school more than 30 years taking Intermediate Algebra to enrich their life.

We see from Table 4.9 that the pass rate of high school students enrolled in Intermediate Algebra is about the same as freshmen who graduated from high school 0-5 years ago, which is about 59%. The pass rate decreases to 42.05% for the students graduated from high school 5-10 years ago, and stays constant for students graduated 10-20 years ago, then it increases to 50% for high school graduation years 20-30 years group, and dropped to 21.43% for older people who graduated from high school more than 30 years ago. Because the oldest groups: the students having high school graduation years 20-30 years or $\geqslant$30 years have such small sample sizes (18 or 14, respectively), so the pass rates of these two groups obtained from these very small samples are not very reliable. Overall, there is a pattern that the pass rate decreases as years since high school graduation increases.

4.2.3 Multicollinearity issues and variable selection

Multicollinearity refers to the problems associated when predictor variables are highly intercorrelated in regression models. This violates some of the basic assumptions in regression models. Multicollinearity is considered to be one of the most severe problems in multiple regression models and is often referred to by social modelers as the "familiar curse". Multicollinearity diagnostics measure how much regressors are related to other regressors and how this affects the stability and variance of the regression estimates. For detailed explanations about multicollinearity see Belsley et al. (2005), Stewart (1987), Velleman and Welsch (1981) and others.

Signs of multicollinearity in a regression analysis include:

1. Large standard errors on the regression coefficient, so that estimates of the true model parameters become unstable.
2. The parameter estimates vary greatly from sample to sample for the same parameters.
3. Drastic changes in the regression estimates after only minor data changes.
4. Unreasonable conclusions reached from the usual tests of significance (such as the wrong

Years since HS graduation	<0	0-5	5-10	10-20	20-30	⩾30	Total
Pass	7	699	37	19	9	3	744
Percentage 1	0.54	51.15	2.83	1.45	0.69	0.23	56.88
Percentage 2	0.94	89.92	4.97	2.55	1.21	0.40	100%
Pass rate	**58.33**	**59.15**	**42.05**	**42.22**	**50.00**	**21.43**	
Not Pass	5	462	51	26	9	11	564
Percentage 3	0.38	35.32	3.90	1.99	0.69	0.84	43.12
Percentage 4	0.89	81.91	9.04	4.61	1.60	1.95	100%
Failing rate	41.67	40.85	57.95	57.78	50.00	78.57	
Total	12	1131	88	45	18	14	1308
Percentage 5	0.92	86.47	6.73	3.44	1.38	1.07	100.0

Table 4.9: Pass rate of students with different years since high school graduation. The first row:"Years since HS graduation" displays the 6 groups we divided according to students' years since high school graduation; The second row, which starts with "Pass", gives the number of students passed the course in each group. "Percentage 1" represents the percentage of passed students in each group over all the 1308 students in the course; "Percentage 2" is the percentage of passed students in each years group over number of passed students (744). The row starting with "Pass rate" provides the percentage of passed students over the total number of students in each of the 6 groups in the course. Similarly, the row "Not pass" provides the number of students who did not pass the course; "Percentage 3" gives the number of failed students in each group over the total number of students in the course (1308), and "Percentage 4" represents the number of failed students in each year group over the number of failed students. "Failing rate" represents the percentage of failed students in each group among the students in each group. "Total" tells the number of students in each group; "Percentage 5" gives the percentage of students in each "Years since HS graduation" group over all students in the course (1308).

sign for a parameter).

5. Extreme correlations between pairs of variables.

6. Omitting a variable from the equation results in smaller regression standard errors.

7. A good fitting model does not give good outside predictions.

We begin with assessing the pairwise correlations of the continuous predictor variables. A Pearson correlation coefficients matrix is displayed in Table 4.10. A typical first approach for

detecting multicollinearity is to inspect the correlation matrix for high pairwise correlations. However, this is not sufficient, since multicollinearity can exist with no pairwise correlations being high. In our case, there is no extreme correlation between two predictor variables identified, and the largest correlation is that between *HS grad. years* and *SAT/ACT* : -0.29, which is not considered big enough to concern for multicollinearity.

Variable	Semester GPA	Course load	SAT/ACT	HS grad. years
Semester GPA	1.000	0.013572	0.1018	−0.00036
Course load	0.13572	1.000	0.00065	−0.12652
SAT/ACT	0.10178	0.00065	1.0000	−0.28692
HS grad. years	−0.00036	−0.12652	−0.28695	1.000

Table 4.10: Multicollinearity dianostics of quantitative predictor variables of the Fall data

Three other commonly used diagnostics for testing multicollinearity are VIF (Variance Inflation Factor), tolerance and condition index. The tolerance is the proportion of variance in a given predictor that is not explained by all of the other predictors, while the VIF is simply 1 / tolerance. The VIF represents a factor by which the variance of the estimated coefficient is multiplied due to the multicollinearity in the model. A VIF greater than 10 is a sign of multicollinearity, while Allison (2012) tends to get concerned when a VIF is greater than 2.50. The variance inflation factors given in Table 4.11 are all small, and the largest is 2.11836. Thus, there is no sign of multicollinearity. We obtained the condition index for each predictor variable and they also suggest that there is no mulitcollinearity problem. We omitted the condition index table for brevity.

Taking into account the results from stairwise elimination for variable selection, the interaction terms between the predictor variables do not appear to be significant. We will use these predictor variables listed in Table 4.11 to fit the models.

Variable	Tolerance	VIF
Intercept		0
Semester GPA	0.95902	1.04273
Course load	0.95667	1.04529
SAT/ACT	0.87359	1.1447
HS grad. years	0.84635	1.18154
High school	0.9612	1.04037
Instructor	0.9731	1.02765
Asian	0.87165	1.14725
White	0.47264	2.11576
Hispanic	0.47206	2.11836

Table 4.11: Collinearity dianostics based on tolerance and variance inflation factor by each predictor variable

4.3 Limitation of Using Multilevel Logistic Models

Based on the description of the data in Section 4.2, one option for modeling the partially ordered data such as students' letter grades is a multilevel logistic model (Wong and Mason, 1985; Gilmour et al., 1985; Gelman and Hill, 2007), in which,

$$\begin{aligned} Y_i|p_i &\overset{ind}{\sim} Bern(p_i) \\ logit(p_i) &= \boldsymbol{X}_i{}'\boldsymbol{\beta} + \boldsymbol{Z}_i{}'\boldsymbol{\tau} \quad i = 1, 2, \ldots, n. \\ \tau_j|\sigma_\tau^2 &\overset{iid}{\sim} N(0, \sigma_\tau^2) \qquad j = 1, 2, \ldots, J. \end{aligned}$$

Y_i is a binary response variable. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are fixed and random design vectors for student i; $\boldsymbol{X}_i$ contains the student's demographic and achievement information and $\boldsymbol{Z}_i$ indicates which instructor the student had. $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of unknown fixed effects and $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_J)'$ is the vector of unknown random effects. σ_τ^2 is the variance of the random effect $\boldsymbol{\tau}$, and n is the number of students, while p is the number of covariates and J is the number of instructors. A probit model can be used by simply replacing the logit function with the inverse standard normal cdf.

With this model, we are forced to make our response variable, the letter grade, into a

binary variable (pass or fail). This means that the passing letter grades ($A+, A, A-, B+, B, B-, C+, C, CR$) are treated equally. Similarly, the failing letter grades ($C-, D+, D, D-, F, NC$) are also treated equally. It is important to note that this model can not describe student success precisely. Suppose there are two classes of 20 students each. In one class, 10 students passed the course with an A and 10 failed with an F, while 10 students in the other class passed the course with a C and 10 failed with a F. The pass rate is the same between the two classes, which is 50%, but the student performance is actually different. Another issue in our data is that many students in the Redesign (the Redesigned course, 54.6%) received an Incomplete at the end of the term. How should we interpret these Incompletes: as a passing letter grade? As a failing letter grade? These students had not yet completed all of the course topics, but they were still working. Both options are extreme because this "go at your own pace" course can be completed over three semesters; there is simply not enough information available to make this determination *a priori*. Many authors forced the letter grades "I", "W", "AUD" to be failing grades to fit a logistic model, while this distorted the information conveyed in the data. An alternative practice is to leave these students out of the study, but eliminating about 70% students with "I","W" or "AUD" in the Redesign would be highly problematic.

4.4 Limitation of Fitting Ordered Probit/Logit Models or Unordered Multinomial Probit/Logit Models

Another option for modeling this type of data is to use an ordered probit/logit model or unordered multinomial probit/logit model. These models enable us to model ordinally scaled or unordered nominal dependent variables with independent variables. For a detailed description of ordered probit or logit models see Section 3.1.3, or McKelvey and Zavoina (1975), McCullagh (1980), Winship and Mare (1984), and Kropko (2007).

An ordered logit model is an extension of the above binary logistic model, but the response data can have more than two ordered categories. Ordered logit or probit models are widely used in economics, marketing and psychology. With this model, the dependent variable has to include ordered cases with about equal distance between neighboring categories. There is a single, unobservable, continuous variable related to this ordered scale, so we can assume a latent variable for the model. We reviewed ordered probit and logit models in Section 3.1.3. Since in our case, some of our response data are ordered (A, B, C, D, with $+/-$s, and F), but others are not ordered and overlap (CR, NC, W, I, AUD). Note that NC, W, I are not necessarily better than D or F, and CR is not necessarily better than A, B, or C. Thus, only part of our data qualifies for the ordered logit models, and we would have to remove observations with responses CR, NC, W, I, AUD to use these models. This would result in a loss of information. Again, a probit version of this model can be easily considered by changing the link function.

Unordered multinomial logit/probit models fit the case that the response categories are mutually exclusive, collectively exhaustive, and without a logical ordering or natural ranking, see Section 3.1.3 for more explanation. Examples for unordered multinormial logit/probit models such as the choice of existent political parties, religion, neighborhood, ways of transportation, etc. Since those categories are unordered, and the probability of falling in each category versus the baseline category is given if we apply this model with several kinds of software like $SAS, R, Stata$ and others. The result can be interpreted as in comparison to the baseline category, each unit increase in the independent variable increase/decrease the probability of selecting (falling in) an alternative category. Obviously, this is not what we want with our data to answer our research questions. Furthermore, part of our responses (letter grades A, B, C, D, with +/-s, and F) are ordered, unordered multinomial logit/probit models discard this natural ordering conveyed by the letter grades and ignore the known relationship. Borooah (2002) stated that if the response variable is ordered but we treat it as unordered and fail to impose a legitimate ranking on the outcome, this might lead to a loss of efficiency. Thus, an unordered multinomial model is not a good fit either.

4.5 A Description of Our Proposed Models

In this section, we propose two new models: the Bayesian partially ordered probit and logit models, which can accommodate data having some ordering but are not fully ordered. Assume that each student has a latent score, S_i, an unknown numerical score for student i. The partially ordered probit model is formulated as:

$$S_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{\tau} + \epsilon_i \tag{4.1}$$

With

$$\begin{aligned} \epsilon_i|\sigma_\epsilon^2 &\overset{iid}{\sim} N(0, \sigma_\epsilon^2) \\ \tau_j|\sigma_\tau^2 &\overset{iid}{\sim} N(\mu_\tau, \sigma_\tau^2) \end{aligned}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, J$. Where, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the fixed and random design vectors, respectively. $\boldsymbol{X}_i$ includes the covariates about student information, and we use $\boldsymbol{Z}_i$ to represent the instructor who taught student i to model the Fall 2012 data. $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of unknown fixed effects, and $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_J)'$ is the vector of unknown random effects. We denote the sample size by n, the number of fixed effects by p, and the number of instructors by J. If the students' exam or course score S_i is known, a reasonable model would be a hierarchical linear model having the same form as Equation 4.1. Our model is much simpler than the item response poset models reviewed in Section 3.1.4,

In our example, we do not know S_i. However, the observed letter grade Y_i gives us information about the latent score S_i. For example, if student i passed, we know that $S_i > 73$ if 73 is the cut-off point for passing the course.

Albert and Chib (1993) examined the situation where Y_i is either a pass or fail with a Bayesian probit model. Note that, if we assume $\epsilon_i \overset{iid}{\sim} N(0, 1)$

$$P(S_i > 73|\boldsymbol{\beta}, \boldsymbol{\tau}) = 1 - \Phi(73 - X_i\beta - Z_i\tau) = \Phi(-(73 - X_i\beta - Z_i\tau)).$$

Which gives us a probit-style probability. For the case of ordered Y_i (for example A, B, C, D with+/- and F), Albert and Chib (1993) assumed a latent continuous random variable distributed as $N(\boldsymbol{X}\boldsymbol{\beta}, 1)$ (see Section 3.1.1), and the observed categorical response $Y_i = \text{j}$ if $\gamma_{j-1} < S_i \leq \gamma_j$, where the bin boundaries $\gamma_1, ..., \gamma_{J-1}$ are typically unknown, and $\gamma_0 \equiv -\infty, \gamma_J \equiv \infty$. Thus, $Si|\beta, \gamma, Y$ follows a truncated normal distribution. This is the standard formulation of a Bayesian ordered probit model.

In our case, including CR, NC, I, W, AUD as possible letter grades, we can only partially order the responses. However, each possible grade does yield a range of possible values of S_i. For example, we can view $Y_i = A+ \Rightarrow 97 \leq S_i, \quad Y_i = A- \Rightarrow 93 \leq S_i < 97, \ldots Y_i = F \Rightarrow 0 \leq S_i < 60$. For the unorderable grades, we assume that $Y_i = CR \Rightarrow 73 \leq S_i$, $Y_i = NC \Rightarrow 0 \leq S_i < 73$, $Y_i = W \Rightarrow 0 \leq S_i < 73$, $Y_i = I \Rightarrow 0 \leq S_i \leq 100$, and $Y_i = AUD \Rightarrow 0 \leq S_i \leq 100$. This assumes that we have little knowledge what the course score of a student receiving an I (Incomplete), or AUD (Audit) was at the time he or she left the course.

The key purpose of this formulation is that we know the conditional distribution of the latent data S_i, conditional on the parameters $(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2)$ and the letter grade (Y_i), has a truncated normal distribution. Symbolically,

$$S_i|\boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\tau}, Y_i, \sigma_\tau^2 \overset{ind}{\sim} TN(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau}, \sigma_\epsilon^2, L_{Y_i}, U_{Y_i}).$$

Let L_{Y_i} denote the lower bound and U_{Y_i} denote the upper bound of S_i given student i received letter grade Y_i. Then, S_i is truncated on the boundaries of each letter scale, and we know that $S_i \in (L_{Y_i}, U_{Y_i})$. Table 4.12 displays the boundaries of course scores corresponding to each letter grade. We used 73 as the cutoff passing score. If a student earned an A+, his course score is between 97 and 100, an A is between 93 and 97, ..., and an F is between 0 and 60. A C- is not considered a passing grade for this course. The score for CR is between 73 and 100, while NC is between 0 and 73. A "W" is treated as a failure in this dissertation. "I" and "AUD" give us the least information about students' course score, thus "I" and "AUD" can take any value between 0 and 100. Other universities may have different

rules concerning non-standard grades. For example, Incomplete may only be awarded to students currently failing. The boundaries for the latent score can be easily modified for their particular institution.

Letter Grade (Y_i)	L_{Y_i}	U_{Y_i}	Letter Grade	L_{Y_i}	U_{Y_i}
A+	97	100	D+	67	70
A	93	97	D	63	67
A-	90	93	D-	60	63
B+	87	90	F	0	60
B	83	87	CR	73	100
B-	80	83	NC	0	73
C+	77	80	W	0	73
C	73	77	I	0	100
C-	70	73	AUD	0	100

Table 4.12: The range of course scores for each letter grade. The range of course scores corresponding to each letter grade Y_i, L_{Y_i} is the lower limit score of a letter grade, and U_{Y_i} is the upper limit of the letter grade.

Because the boundaries of the latent score S_i, given the letter grade Y_i are known, as displayed in Table 4.12, we are able to allow the variance of the error terms σ_ϵ^2 to be variable while the parameters remain identifiable. We include the random effect in the model because the grades of students taught by the same instructor may be correlated. Ignoring this nested feature of the data will cause underestimating the variance of the estimated coefficients, and resulting in inconsistent parameter estimates as well, if the underlying relationship between the outcome variable and the explanatory variables is nonlinear (Rodriguez and Goldman, 1995). The variance of the random effects $\boldsymbol{\tau}$ is σ_τ^2.

The key point to take away from this model is that S_i is unknown, and we only observe the partially ordered categorical variable Y_i. However, the letter grade that a student receives does provide information on the possible value of S_i. With this latent variable representation, we can not only model ordered and unordered multinomial response data, but also partially

ordered categorical outcomes. The knowledge from the partial ordering is also preserved. Specifically, our model has exactly the form of a linear mixed model with random effects. The latent score of a student taught by any instructor listed can be predicted if the students' information is known.

4.5.1 Bayesian formulation of partially ordered probit model with random effect

The latent variable formulation of the probit and logit models described earlier naturally lead themselves to Bayesian models. Sampling from the posterior allows us to make predictions while taking into account parameter uncertainty. Bayesian analysis can easily incorporate truncated distributions and also allows us to use subjective information in our priors.

To estimate the parameters $\boldsymbol{\beta}, \boldsymbol{\tau}$, σ_τ^2 and σ_ϵ^2, we use Bayesian inference with a prior for each parameter represented by $\pi(\boldsymbol{\beta}), \pi(\boldsymbol{\tau}), \pi(\sigma_\tau^2)$, and $\pi(\sigma_\epsilon^2)$, respectively. The likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2 | \boldsymbol{Y}) = \Pr(\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2) = \prod_{i=1}^{n} \Pr(Y_i | \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2),$$

and posterior distribution

$$\pi(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2 | \boldsymbol{Y}) = \int_{\mathbf{S}} \pi(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2, \boldsymbol{S} | \boldsymbol{Y}) d\boldsymbol{S}$$

$$= \int_{\mathbf{S}} \frac{L(\boldsymbol{Y} | \boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2) \pi(\boldsymbol{S} | \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2) \pi(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2)}{m(\boldsymbol{Y})} d\boldsymbol{S},$$

where $L(\boldsymbol{Y} | \boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2) \pi(\boldsymbol{S} | \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2)$ is the joint likelihood of $\boldsymbol{Y}$ and $\boldsymbol{S}$, $m(\boldsymbol{Y})$ is the marginal distribution of $\boldsymbol{Y}$,

$$\begin{aligned} m(\boldsymbol{Y}) \;\; = \;\; & \int_{\mathbf{S}} \int_{\boldsymbol{\beta}} \int_{\boldsymbol{\tau}} \int_{\sigma_\tau^2} \int_{\sigma_\epsilon^2} L(\boldsymbol{Y} | \boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2) \pi(\boldsymbol{S} | \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2) \\ & \pi(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2) d\boldsymbol{\beta} d\boldsymbol{\tau} d\sigma_\tau^2 d\sigma_\epsilon^2 d\boldsymbol{S}. \end{aligned}$$

The multidimensional integral looks intimidating (especially when there are many $\boldsymbol{\beta}$s and $\boldsymbol{\tau}$s), and obviously it is not simple to find. Following Albert and Chib (1993), we defined

a Gibbs sampler utilizing the latent variable $\boldsymbol{S}$, that is, the sample we obtained by sampling from full conditional distributions successively with Gibbs sampling converges to the posterior distribution, see Section 3.2.2 for a description of Gibbs sampling.

Prior Distribution

To complete a Bayesian analysis, we need prior distributions for $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, μ_τ (the mean of random effects $\boldsymbol{\tau}$s), σ_τ^2s, and σ_ϵ^2. For computational convenience, we assumed an improper uniform prior for $\boldsymbol{\beta}$ (we had little prior knowledge about $\boldsymbol{\beta}$), a normal prior for the instructor random effects $\boldsymbol{\tau}$ (conditional on hyperparameters), an improper uniform prior for the average instructor random effect μ_τ, and inverse gamma priors for the variance of the instructor random effects σ_τ^2 and the latent score error variance σ_ϵ^2. Symbolically:

$$\begin{aligned}
\pi(\boldsymbol{\beta}) &\propto 1, \\
\tau_j|\mu_\tau, \sigma_\tau^2 &\overset{iid}{\sim} N(\mu_\tau, \sigma_\tau^2) \quad \text{for } j = 1, 2, \ldots, J, \\
\pi(\mu_\tau) &\propto 1, \\
\sigma_\tau^2 &\overset{iid}{\sim} IG(\alpha_\tau, \beta_\tau), \\
\sigma_\epsilon^2 &\overset{iid}{\sim} IG(\alpha_\epsilon, \beta_\epsilon),
\end{aligned}$$

where, $\alpha_\tau, \beta_\tau, \alpha_\epsilon, \beta_\epsilon$ are fixed constants. In practice, these may be chosen to reflect subjective knowledge or prior ignorance. Typical conditional independence is assumed, meaning:

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2) &= \pi(\boldsymbol{\beta}|\boldsymbol{\tau}, \mu_\tau, \sigma_\tau^2, \sigma_\epsilon^2)\pi(\boldsymbol{\tau}|\mu_\tau, \sigma_\tau^2, \sigma_\epsilon^2)\pi(\mu_\tau|\sigma_\tau^2, \sigma_\epsilon^2)\pi(\sigma_\tau^2|\sigma_\epsilon^2)\pi(\sigma_\epsilon^2) \\
&= \pi(\boldsymbol{\beta})\pi(\boldsymbol{\tau}|\mu_\tau, \sigma_\tau^2)\pi(\mu_\tau)\pi(\sigma_\tau^2)\pi(\sigma_\epsilon^2)
\end{aligned}$$

for all unobservables. In the examples considered in this work, we set $a_\tau = a_\epsilon = 300$ and $b_\tau = b_\epsilon = 6$. This was based on our previous experience with course grades. Other applicants of this model may require modifications to this prior formulation. Specifically, a user may assume subjective priors for $\boldsymbol{\beta}$ and μ_τ if prior knowledge is available.

These priors imply that our posterior distribution including $\boldsymbol{S}$ satisfies:

$$\pi(\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\epsilon^2, \sigma_\tau^2 | \boldsymbol{Y}) \quad \propto \quad f(\boldsymbol{S}|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\epsilon^2)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\tau}|\mu_\tau, \sigma_\tau^2)\pi(\mu_\tau)\pi(\sigma_\epsilon^2)\pi(\sigma_\tau^2).$$

The Disribution of Parameters in the Partially Ordered Multinomial Probit Models

In this section, we develop the partially ordered probit model stated in Equation 4.1 step by step. We start with simpler models, and gradually add complexity. The partially ordered probit model encapsulates the following several simpler models, and each model can be applied to partially ordered categorical response data with the algorithm stated in Section 4.5 and 4.6. Which model to choose depends on the specific form of the data and personal preference.

Model 1: Partially ordered probit model with only fixed effects. We start from the simplest: the partially ordered multinomial probit model with only fixed effects. Albert and Chib (1993) first applied this model with a Bayesian approach on ordered polychotomous response data, and we generalized to partially ordered response data. The model is expressed in Equation 4.2, and this is the standard latent structure model:

$$S_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \tag{4.2}$$

$$\epsilon_i|\sigma_\epsilon^2 \quad \overset{iid}{\sim} \quad N(0, 1),$$

for $i = 1, 2, \ldots, n$.

Using a noninformative prior for $\boldsymbol{\beta}$, and fixing the variance of S_i at 1, S_i is distributed

as $N(\boldsymbol{X}_i\boldsymbol{\beta}, 1)$. The conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{S}$ and $\boldsymbol{Y}$ is:

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\boldsymbol{S},\boldsymbol{Y}) &\propto \pi(\boldsymbol{Y}|\boldsymbol{S},\boldsymbol{\beta})\pi(\boldsymbol{S}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \\
&\propto L(\boldsymbol{S}|\boldsymbol{\beta}) \\
&\propto exp\left[-\frac{1}{2}(\boldsymbol{S}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{S}-\boldsymbol{X}\boldsymbol{\beta})\right] \\
&\propto exp\left[-\frac{1}{2}(\boldsymbol{S}^\top\boldsymbol{S}-\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{S}-\boldsymbol{S}^\top\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})\right] \\
&\propto exp\left[-\frac{1}{2}(\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}-2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{S}+\boldsymbol{S}^\top\boldsymbol{X}^\top(\boldsymbol{X}^\top\boldsymbol{X})^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{S})\right] \\
&\propto exp\left[-\frac{1}{2}(\boldsymbol{\beta}-(\boldsymbol{X}^\top\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^\top\boldsymbol{S})^\top(\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{\beta}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{S})\right].
\end{aligned}$$

which is the kernel of a multivariate normal distribution with mean $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{S}$ and covariance matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. This means that $\boldsymbol{\beta}$ conditional on $\boldsymbol{S}$ and $\boldsymbol{Y}$ is distributed as: $\boldsymbol{\beta}|\boldsymbol{S},\boldsymbol{Y} \sim N((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{S}, \quad (\boldsymbol{X}^\top\boldsymbol{X})^{-1})$.

Model 2: Partially ordered probit model with fixed and random effects Now, we add the random effect $\boldsymbol{\tau}$ to the model in Equation 4.2 to describe the nested feature of the data. The model can be expressed as:

$$S_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau} + \epsilon_i, \tag{4.3}$$

$$\begin{aligned}
\epsilon_i &\overset{iid}{\sim} N(0, \sigma_\epsilon^2 = 1), \\
\tau_j &\overset{iid}{\sim} N(0, \sigma_\tau^2),
\end{aligned}$$

with the noninformative prior for $\boldsymbol{\beta}$. Again, the variance σ_ϵ^2 is fixed at 1. The conditional distribution of the latent score, $\boldsymbol{S}$, given other parameters is: $S|\boldsymbol{\beta},\boldsymbol{\tau} \sim (\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\tau}, 1)$. Using matrix expression for all parameters, the conditional distribution of $\boldsymbol{\beta}$ given other parameters

is:

$$
\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{S},\boldsymbol{\tau} &\propto L(\boldsymbol{S}|\boldsymbol{\beta},\boldsymbol{\tau})\pi(\boldsymbol{\beta}) \\
&\propto L(\boldsymbol{S}|\boldsymbol{\beta},\boldsymbol{\tau}) \\
&\propto exp\bigg[-\frac{1}{2}(\boldsymbol{S}-\boldsymbol{X\beta}-\boldsymbol{Z\tau})^{\top}((\boldsymbol{S}-\boldsymbol{X\beta}-\boldsymbol{Z\tau})\bigg] \\
&\propto exp\bigg[-\frac{1}{2}(-(\boldsymbol{S})^{\top}\boldsymbol{X\beta}-\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{S}+\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X\beta}+\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{Z\tau}+\boldsymbol{\tau}\boldsymbol{Z}^{\top}\boldsymbol{X\beta})\bigg] \\
&\propto exp\bigg[-\frac{1}{2}(\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X\beta}-2\boldsymbol{\beta}^{\top}(\boldsymbol{X}^{\top}\boldsymbol{S}-\boldsymbol{X}^{\top}\boldsymbol{Z\tau})+\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{X\beta})\bigg] \\
&\propto exp\{-\frac{1}{2}\big[\boldsymbol{\beta}-(\boldsymbol{X}^{\top}\boldsymbol{X})^{-\mathbf{1}}(\boldsymbol{X}^{\top}\boldsymbol{S}-\boldsymbol{X}^{\top}\boldsymbol{Z}\tau)\big]^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}\left[\boldsymbol{\beta}-(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}(\boldsymbol{X}^{\top}\boldsymbol{S}-\boldsymbol{X}^{\top}\boldsymbol{Z\tau})\right]\}
\end{aligned}
$$

We recognize that this expression is the kernel of normal distribution. Hence the conditional distribution of $\boldsymbol{\beta}$ given all other parameters is normal:
$\boldsymbol{\beta}|\boldsymbol{S},\boldsymbol{\tau} \sim N\left((\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}(\boldsymbol{X}^{\top}\boldsymbol{S}-\boldsymbol{X}^{\top}\boldsymbol{Z\tau}),(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\right)$.

The conditional distribution for the random effect $\boldsymbol{\tau}$ in model 2 Applying Model 2 to the course redesign evaluation at the University, $\boldsymbol{\tau}$ is the effect of different instructor on student success measured by the letter grades. Let n_j represents the number of students with instructor j, and student's score is normally distributed: $S_i|\boldsymbol{\beta},\boldsymbol{\tau},\sigma_{\tau}^2,\sigma_{\epsilon}^2 \sim N(\boldsymbol{X}_i'\boldsymbol{\beta}+\boldsymbol{Z}_i'\boldsymbol{\tau},1)$, the variance of the error terms is still fixed. Recall that the prior of τ_j is

$$\tau_j|\mu_{\tau},\sigma_{\tau}^2 \overset{iid}{\sim} N(u_{\tau},\sigma^2\tau) \quad \text{for } j = 1, 2, \ldots, \text{J.}$$

The conditional distribution of τ_j given $\boldsymbol{S}$ and $\boldsymbol{\beta}$ is:

$$
\begin{aligned}
& \tau_j|\boldsymbol{S},\boldsymbol{\beta},\mu_\tau,\sigma_\tau^2 \propto L(\boldsymbol{S}|\boldsymbol{\beta},\boldsymbol{\tau})\pi(\boldsymbol{\tau}), \\
\propto & \prod_{i=n_{j-1}+1}^{n_j} \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}(S_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{\tau}_j)^2\right] \quad \times \frac{1}{\sqrt{2\pi\sigma_\tau^2}} exp\left[-\frac{1}{2\sigma_\tau^2}(\boldsymbol{\tau}_j - \mu_\tau)^2\right], \\
\propto & \; exp\left\{-\frac{1}{2}\left[\sum_{i=n_{j-1}+1}^{n_j}(\boldsymbol{S}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{\tau}_j)^2 + \frac{(\boldsymbol{\tau}_j - \mu_\tau)^2}{\sigma_\tau^2}\right]\right\}, \\
\propto & \; exp\left\{-\frac{1}{2}\left[\sum_{i=n_{j-1}+1}^{n_j}(S_i - \boldsymbol{X}_i\boldsymbol{\beta})^2 - 2\sum_{i=1}^{n_j}(S_i - \boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{\tau}_j + n_j\tau_j^2 + \frac{1}{\sigma_\tau^2}(\tau_j^2 - 2\boldsymbol{u}_\tau\tau_j + \mu_\tau^2)\right]\right\}, \\
\propto & \; exp\left\{-\frac{1}{2}\left[\tau_j^2(n_j + \frac{1}{\sigma_\tau^2}) - 2\tau_j\left(\frac{\mu_\tau}{\sigma_\tau^2} + \sum_{i=1}^{n_j}(S_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)\right]\right\}, \\
\propto & \; exp\left\{-\frac{n_j + \frac{1}{\sigma_\tau^2}}{2}\left[\tau_j^2 - 2\tau_j(n_j + \frac{1}{\sigma_\tau^2})^{-1}\left(\frac{\mu_\tau}{\sigma_\tau^2} + \sum_{i=1}^{n_j}(S_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)\right]\right\}, \\
\propto & \; exp\left\{-\frac{1}{2(n_j + \frac{1}{\sigma_\tau^2})^{-1}}\left[\tau_j^2 - 2\tau_j(n_j + \frac{1}{\sigma_\tau^2})^{-1}\left(\frac{\mu_\tau}{\sigma_\tau^2} + \sum_{i=1}^{n_j}(S_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)\right]\right\},
\end{aligned}
$$

with $n_0 = 0$. Thus, τ_j is normally distributed with mean $=$
$(n_j + \frac{1}{\sigma_\tau^2})^{-1}\left(\frac{\mu_\tau}{\sigma_\tau^2} + \sum_{i=1}^{n_j}(S_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)$ and variance $= (n_j + \frac{1}{\sigma_\tau^2})^{-1}$. Symbolically,

$$
\tau_j|\boldsymbol{S},\boldsymbol{\beta} \sim N\left((n_j + \frac{1}{\sigma_\tau^2})^{-1} \times \left(\frac{\mu_\tau}{\sigma_\tau^2} + \sum_{i=1}^{n_j}(\boldsymbol{S}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right), \quad (n_j + \frac{1}{\sigma_\tau^2})^{-1}\right)
$$

The conditional distribution for σ_τ^2 in Model 2 To derive the conditional distribution for σ_τ^2, we need a prior. An Inverse Gamma prior for σ_τ^2, $\pi(\sigma_\tau^2) \propto Inverse\ Gamma(\alpha_\tau, \beta_\tau)$, makes the conditional distribution of $\sigma_\tau^2|\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{S}$ conjugate. It is reasonable since we can adjust the value of α_τ, β_τ according to our prior knowledge on students' course score variation between instructors. The prior for σ_τ^2 is:

$$
\pi(\sigma_\tau^2|\alpha_\tau,\beta_\tau) = \frac{{\beta_\tau}^{\alpha_\tau}}{\Gamma(\alpha_\tau)}(\sigma_\tau^2)^{-\alpha_\tau-1}exp^{(\frac{-\beta_\tau}{\sigma_\tau^2})}
$$

with mean $= \frac{\beta_\tau}{\alpha_\tau - 1}$ for $\alpha_\tau > 1$ and variance $= \frac{{\beta_\tau}^2}{(\alpha_\tau-1)^2(\alpha_\tau-2)}$ for $\alpha_\tau > 2$. Then the conditional distribution for $\sigma_\tau^2|\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}$ is:

$$\begin{aligned}
\sigma_\tau^2|\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau} &\propto L(\boldsymbol{S}|\boldsymbol{\beta}, \boldsymbol{\tau})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\tau}|\sigma_\tau^2)\pi(\sigma_\tau^2) \\
&\propto \pi(\boldsymbol{\tau}|\sigma_\tau^2)\pi(\sigma_\tau^2) \\
&\propto \left(\prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi\sigma_\tau^2}} exp\left[-\frac{1}{2\sigma_\tau^2}(\tau_j - \mu_\tau)^2\right]\right) \quad \times \frac{{\beta_\tau}^{\alpha_\tau}}{\Gamma(\alpha_\tau)}(\sigma_\tau^2)^{-\alpha_\tau - 1} exp(\frac{-\beta_\tau}{\sigma_\tau^2}) \\
&\propto (\frac{1}{\sigma_\tau^2})^{\frac{J}{2}+\alpha_\tau+1} exp\left(-\frac{1}{\sigma_\tau^2}\left[\beta_\tau + \frac{\sum_{j=1}^J(\tau_j - \mu_\tau)^2}{2}\right]\right) \\
&\propto (\sigma_\tau^2)^{-(\frac{J}{2}+\alpha_\tau)-1} exp\left(-\frac{1}{\sigma_\tau^2}\left[\beta_\tau + \frac{\sum_{j=1}^J(\tau_j - \mu_\tau)^2}{2}\right]\right)
\end{aligned}$$

which is the kernel of Inverse Gamma distribution with $\alpha_{\sigma_\tau^2} = \frac{J}{2} + \alpha_\tau$ and $\beta_{\sigma_\tau^2} = \beta_\tau + \frac{\sum_{j=1}^J(\tau_j - u_\tau)^2}{2}$. Thus the conditional distribution of $\sigma_\tau^2|\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}$ is:

$$\sigma_\tau^2|\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau} \sim IG(\frac{J}{2} + \alpha_\tau,\ \beta_\tau + \frac{\sum_{j=1}^J(\tau_j - u_\tau)^2}{2})$$

The mean of τ_j can be fixed at 0, or allowed to be random. If we would like to make μ_τ a random variable, then use a flat prior $\pi(\mu_\tau) \propto 1$, the distribution of $\mu_\tau|\tau_j, \sigma_\tau^2$ will be normal with mean $\bar{\tau} = \frac{1}{J}\sum \tau_j$, and variance $\frac{\sigma_\tau^2}{J}$, symbolically represented by $\mu_\tau|\tau_j, \sigma_\tau^2 \sim N(\bar{\tau}, \frac{\sigma_\tau^2}{J})$.

So far, the variance of $\boldsymbol{S}_i|\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2$ has been made fixed as 1, and this was done for identifiability concerns. We may allow it to vary if the boundaries of $\boldsymbol{S}$ for each partially ordered categories are known (as in our case).

The conditional distribution of σ_ϵ^2 in model 2 With σ_ϵ^2 being random, the model becomes:

$$S_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau} + \epsilon_i, \tag{4.4}$$

$$\begin{aligned}
\epsilon_i &\overset{iid}{\sim} N(0, \sigma_\epsilon^2), \\
\tau_j &\overset{iid}{\sim} N(0, \sigma_\tau^2).
\end{aligned}$$

We would like to use a subjective conjugate Inverse Gamma prior for σ_ϵ^2 to ensure the conjugacy of the full conditional distribution, and incorporate prior knowledge about variability of students' course score. The prior of σ_ϵ^2 is defined as:

$$\pi(\sigma_\epsilon^2|\alpha_\epsilon, \beta_\epsilon) = \frac{{\beta_\epsilon}^{\alpha_\epsilon}}{\Gamma(\alpha_\epsilon)}(\sigma_\epsilon^2)^{-\alpha_\epsilon-1}exp(\frac{-\beta_\epsilon}{\sigma_\epsilon^2}).$$

Let us begin with the likelihood and prior to derive the full conditional distribution of σ_ϵ^2:

$$\begin{aligned}
\sigma_\epsilon^2|\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \boldsymbol{\epsilon} &\propto L(\boldsymbol{S}|\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\epsilon})\pi(\boldsymbol{\beta})L(\boldsymbol{\tau}|\sigma_\tau^2)\pi(\sigma_\tau^2)L(\boldsymbol{\epsilon}|\sigma_\epsilon^2)\pi(\sigma_\epsilon^2) \\
&\propto L(\boldsymbol{\epsilon}|\sigma_\epsilon^2)\pi(\sigma_\epsilon^2) \\
&\propto \prod_{i=1}^{n}(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}exp\left[\frac{1}{2\sigma_\epsilon^2}(\epsilon_i-\mu_\epsilon)^2\right] \times (\frac{1}{\sigma_\epsilon^2})^{\alpha_\epsilon+1}exp(-\frac{\beta_\epsilon}{\sigma_\epsilon^2}) \\
&\propto \frac{1}{(\sigma_\epsilon^2)^{(1/n)}}exp\sum_{i=1}^{n}\frac{\epsilon_i^2}{2\sigma_\epsilon^2} \times (\frac{1}{\sigma_\epsilon^2})^{\alpha_\epsilon+1}exp(-\frac{\beta_\epsilon}{\sigma_\epsilon^2}) \\
&\propto \left(\frac{1}{\sigma_\epsilon^2}\right)^{(\frac{n}{2}+\alpha_\epsilon+1)}exp\sum_{i=1}^{n}(\frac{\epsilon_i^2}{2\sigma_\epsilon^2}-\frac{\beta_\epsilon}{\sigma_\epsilon^2}) \\
&\propto {\sigma_\epsilon^2}^{-(\frac{n}{2}+\alpha_\epsilon)-1}exp\left[-\frac{1}{\sigma_\epsilon^2}\left(\beta_\epsilon+\sum_{i=1}^{n}\frac{\epsilon_i^2}{2}\right)\right]
\end{aligned}$$

where $\mu_\epsilon = 0$ is the mean of the error terms, and $\epsilon_i = S_i - \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau}$. Thus, the conditional distribution of σ_ϵ^2 given other parameters is also Inverse Gamma: $\sigma_\epsilon^2|\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \boldsymbol{\epsilon} \sim IG(\frac{n}{2} + \alpha_\epsilon,\ \beta_\epsilon + \sum_{i=1}^{n}\frac{\epsilon_i^2}{2})$.

If a logit style model is desired, O'Brien and Dunson (2004) showed that a simple change of the distribution of ϵ_i results in a logit marginal distribution. The logit link can be implemented in the model listed in Equation (4.1) by changing the distribution of $\boldsymbol{\epsilon}$ to $\epsilon_i \overset{ind}{\sim} logistic(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau}, \sigma_\epsilon^2)$. The predicted probability with the logit link is easier to interpret than with the probit link as described in Section 3.1.1.

4.6 Computation

Bayesian inference is centered around the posterior distribution, the distribution of the parameters given observables. Unfortunately, computing the posterior often involves computing high dimensional integrals rarely available in a closed form. Instead, it is often easier to sample from the posterior distribution. Those samples, combined with ergotic theorems, can be used to conduct inference on parameters and make predictions. The Gibbs sampler (Geman and Geman, 1984; Tanner and Wong, 1987), described in Section 3.2.2, is a Markov Chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult to do. Gibbs sampling constructs a Markov Chain which has stationary distribution equal to the target posterior distribution. For an introduction to Gibbs sampling, see Casella and George (1992).

4.6.1 Gibbs Sampling for the Probit Model

Gibbs sampling can work well when it is easy to sample from the full conditional distributions of each unknown parameter (the distribution of that parameter given all other parameters and the data). Iteratively sampling from the full conditionals builds the Markov chain. Following Tanner and Wong (1987) and Albert and Chib (1993), we use the latent score as a parameter to facilitate the Gibbs sampler.

Recall that conditional on the observed letter grade Y_i and all other parameters, S_i follows a truncated normal distribution (the letter grade tells us the range of possible values of the latent score). The conjugacy of this model makes the full conditionals of each parameter known. Generally, the conditional distributions of $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, and μ_τ are normal while σ_τ^2 and σ_ϵ^2 are inverse gamma. The specific full conditional distributions were defined in Section 4.5.1.

A Gibbs sampler for the probit model with random effect is described as follows: at the t-th iteration, each parameter is updated by sampling from the following full conditional

distributions:

$$\begin{aligned}
\boldsymbol{\beta}^{(t)} &\sim \pi(\boldsymbol{\beta}|\boldsymbol{\tau}^{(t-1)}, \mu_\tau^{(t-1)}, \sigma_\tau^{2(t-1)}, \sigma_\epsilon^{2(t-1)}, \boldsymbol{Y}) \\
\boldsymbol{\tau}^{(t)} &\sim \pi(\boldsymbol{\tau}|\boldsymbol{\beta}^{(t)}, \mu_\tau^{(t-1)}, \sigma_\tau^{2(t-1)}, \sigma_\epsilon{}^{2(t-1)}, \boldsymbol{Y}) \\
\mu_\tau^{(t)} &\sim \pi(\mu_\tau|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \sigma_\tau^{2(t-1)}, \sigma_\epsilon^{2(t-1)}, \boldsymbol{Y}) \\
\sigma_\tau^{2(t)} &\sim \pi(\sigma_\tau^2|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_\tau^{(t)}, \sigma_\tau^{2(t-1)}, \sigma_\epsilon^{2(t-1)}, \boldsymbol{Y}) \\
\sigma_\epsilon^{2(t)} &\sim \pi(\sigma_\tau^2|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_\tau^{(t)}, \sigma_\tau^{2(t)}, \boldsymbol{Y}) \\
\boldsymbol{S}^{(t)} &\sim \pi(\boldsymbol{S}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_\tau^{(t)}, \sigma_\tau^{2(t)}, \sigma_\epsilon^{2(t)}, \boldsymbol{Y})
\end{aligned}$$

The elements of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are updated in blocks to improve mixing (Roberts and Sahu, 1997). We assessed convergence by monitoring trace plots and the Gelman-Rubin diagnostic (Gelman and Rubin, 1992). A thinning of the draws was also done, that is, we kept every 10th observation among the draws to reduce autocorrelation among the draws in each parameter sequence and to conserve space.

We used Gibbs sampling to fit Model 2 with the data described in Section 4.2 (the Intermediate Algebra student data in the Fall 2012 at the University). We ran the Gibbs sampler for 1,000,000 iterations. Figure 4.2 shows that the trace plots of the posterior draws of the parameters look acceptable, pretty random points, with little autocorrelation or dependency on the initial values. We have no evidence that the chain has not reached its stationary distribution. These draws may be used to make inferences.

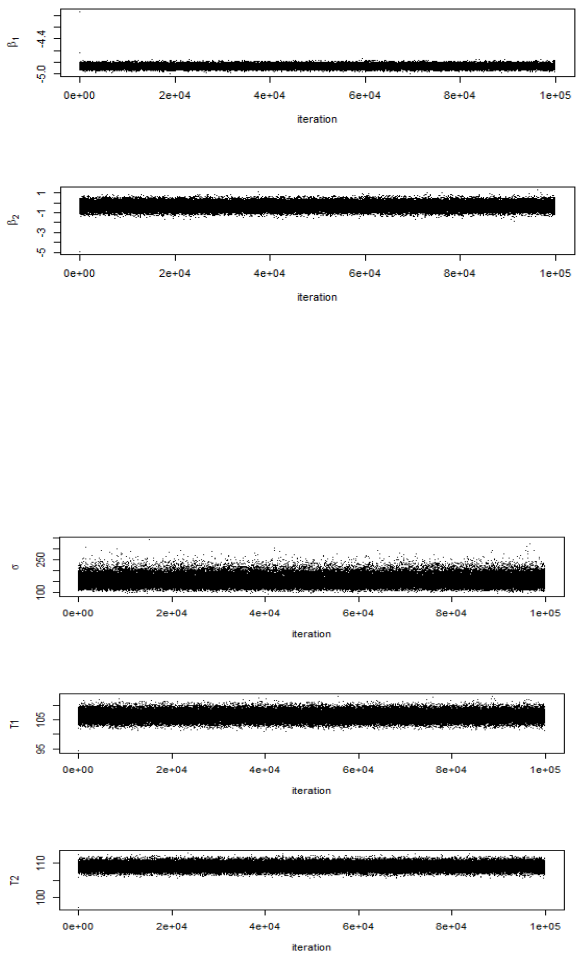

Figure 4.2: Trace plots for $\beta_1(SAT/ACT)$, $\beta_2(Course\ load)$ and σ_ϵ^2 as well as the first two τs. The trace plots for the other parameters are similar.

4.6.2 Importance Sampling for the Logit Model

Recall that in Section 4.5 we described how a logit model may be implemented by changing the distribution of $\boldsymbol{S}_i$ in the model on Equation (4.2) from normal with variance σ_ϵ^2 to a logistic distribution with scale parameter $\sigma^2_{\epsilon_{logit}}$. This model is no longer conditionally conjugate, ruling out using an easy Gibbs sampler for model fitting, but similar to O'Brien and Dunson (2004), we may use importance sampling.

Importance sampling can reduce variance and increase the efficiency of Monte Carlo algorithms for estimating integrals (Ferrenberg and Swendsen, 1988; Geweke, 1989). Instead of drawing from the target distribution, importance sampling samples from an easier "importance" distribution. The ratio of the target distribution's density to the density of the importance distribution is used as weights to estimate expectations of the target distribution. See Section 3.2.2 for a description about importance sampling.

In our example, since we already have sampled from the posterior distribution of the probit model, importance sampling is an easy way to estimate parameters in the partially ordered logit model. Thus, we use the probit model (π_{probit}) as the importance distribution to estimate moments of the logit model (π_{logit}). Let $g(\cdot)$ represent a function of any parameter, and $\boldsymbol{\theta}$ is all the unknown parameters. The importance sampling estimator for the expected value of $g(\boldsymbol{\theta})$, based on T draws from the full posterior of the probit model is :

$$\begin{aligned}
\widehat{\mathrm{E}_{\pi_{logit}}(g(\boldsymbol{\theta}))} &\approx \frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{\theta}^{(t)}) \times \frac{\pi_{logit}(\boldsymbol{\theta}^{(t)}|\boldsymbol{Y})}{\pi_{Probit}(\boldsymbol{\theta}^{(t)}|\boldsymbol{Y})} \\
&\approx \frac{1}{\sum_{t=1}^{T} w_t}\sum_{t=1}^{T} g(\boldsymbol{\theta}^{(t)}) w_t
\end{aligned}$$

With importance weights

$$w_t = \frac{\pi_{logit}(\boldsymbol{\theta}^{(t)}|\boldsymbol{Y})}{\pi_{probit}(\boldsymbol{\theta}^{(t)}|\boldsymbol{Y})}$$

In our case the importance weights are defined as:

$$w_t = \frac{\pi_{logit}(\boldsymbol{S}^{(t)}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_t^{(t)}, \sigma_\epsilon^{2(t)}, \sigma_\tau^{2(t)}, \boldsymbol{Y})}{\pi_{probit}(\boldsymbol{S}^{(t)}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_t^{(t)}, \sigma_\epsilon^{2(t)}, \sigma_\tau^{2(t)}, \boldsymbol{Y})}$$

and $\boldsymbol{\theta}^{(t)}$ is the t-th draw of that particular parameter from the probit model samples. The choice of using the probit model as the importance distribution both because it is practical (we outlined how to obtain these draws in Section 4.6.1) and in our experience it works well. Recall that if there are any huge weights, then some of the estimates under the target distribution will be biased, so the variance of the estimator will be increased. For this purpose, we made a histogram to check the distribution of the weights in Figure 4.3. We see from the plot that there are no extreme values for importance weights. This indicates that the estimator obtained from this importance sampling should work well.

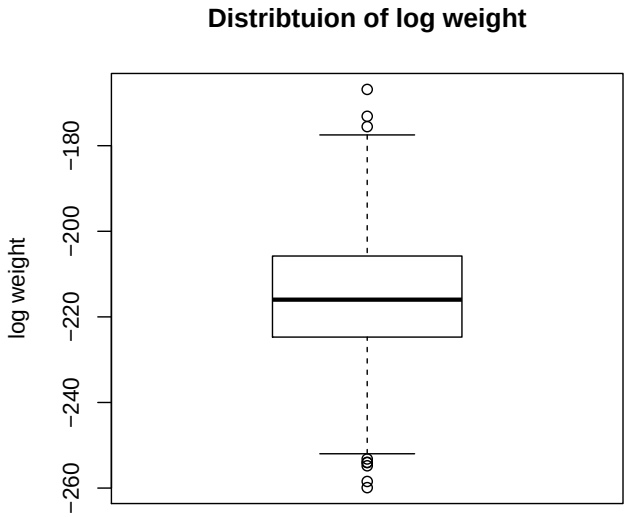

Figure 4.3: Boxplot of the logarithm of importance weights.

The parameter estimates from Gibbs sampling and importance sampling are discussed in Section 4.10.1.

4.7 Gain in Using the Partially Ordered Probit/Logit Models and Extensions

The latent partially ordered representation described above allows us to model all the grading information in the data. That means we do not have to throw away information by forcing the data into a standard binary logistic model or an ordered probit/logit model. With our course redesign evaluation project, not only can students' ordered letter grades A, B, C, D (with $+/-$), F be used in the model, but also other partially ordered grades such as CR, NC, W, I, AUD. We can use all this information to give a more precise prediction of student performance, which is measured by a latent variable, students' course score S_i. The predicted probability of a student passing the course ($Pr(S_i > 73)$) can be estimated based on the demographic and achievement data for a student ($\boldsymbol{X}$) and which instructor taught him or her ($\boldsymbol{Z}$).

With these models, some categories that do not convey much information are treated like missing data that falls within a broad range. For example, the course score for grades "I" and "AUD" can be anything between 0 and 100; we do not know these scores because the students did not finish or their performance in the course was not actually recorded. The course scores of these students are predicted by the model from the data we have about these students.

The latent variable model is easily interpreted for non-statisticians. Introducing the latent variable makes the model take the same form as a linear regression model. As we have reviewed in Section 3.1.1, standard probit and logit models are interpreted as probabilities, it is hard to interpret how the probability changes according to the changing of predictor

variables even for the logit model, do not mention the widely criticized esoteric interpretation of the probit model. Introducing a latent variable, the latent score S_i changes directly in responds to the changing of predictor variables $\boldsymbol{X}$ and $\boldsymbol{Z}$, as with the linear regression. Thus, non-statisticians do not have to have the knowledge about link functions and Bayesian modeling to interpret the models at a high level. Some knowledge in algebra is sufficient to interpret this model.

When only binary grades are observed, the partially ordered probit/logit models collapses to a standard probit/logit model. Similarly, when only ordered grades are observed, the models collapse to an ordered probit/logit model. Thus, the partially ordered framework includes both binary and ordered data as special cases. The partially ordered probit or logit models are flexible to use for either binary, ordered or partially ordered response data.

It is easy to fit the model with Gibbs sampling and importance sampling. Although the joint posterior distribution of parameters $(\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mu_\tau, \sigma_\tau^2, \sigma_\epsilon^2)$ is complicated in the sense that it is difficult to normalize, since our probit model is conditionally conjugate, the Gibbs sampler can be used to fit it. It is computationally easy to sample from the conjugate full conditional distribution, and the algorithm converges fast. Then based on the probit model, importance sampling can be used to estimate the parameters of the logit model. In our experience, estimating parameters and making predictions take less than 30 minutes of computer time.

4.8 Identifiability concern

Indentifiability is always an issue in latent variable models. Unidentifiable models can lead to issues of convergence when attempting to fit the model (Gelfand and Sahu, 1999) and interpretation (Dawid, 1979). As defined by Basu (1983), a parameter θ is identifiable by data X_i if different values of θ yields different distribution of X_i. Bayesian identifiabililty (Dawid, 1979) is concerned with whether observing the data increases our prior knowledge

about the parameter. Gelfand and Sahu (1999) found that trajectories of the Markov chain for components of the parameter will tend to drift to very extreme values, have difficulty converging to the stationary, and can lead to unstable computation if there are unidentifiable parameters.

The existence of an intercept μ_0 in Equation 4.1 would produce an unidentifiable model because the random effects have a non-zero mean, since with an intercept μ_0, the model becomes:

$$\begin{aligned} S_1 &= u_0 + \boldsymbol{X}_i\boldsymbol{\beta} + \tau_1 \\ S_2 &= u_0 + \boldsymbol{X}_i\boldsymbol{\beta} + \tau_2 \\ \dots & \\ S_n &= u_0 + \boldsymbol{X}_i\boldsymbol{\beta} + \tau_J \end{aligned}$$

Where n represents the number of observations, and J is the length of the random effect. The u_0s here are unidentifiable unless $\sum_{i=1}^{n} X_i\boldsymbol{\beta} + \tau_j = 0$, which is unreasonable in practical situations. Hence, we have set the intercept $u_0 = 0$. The convergence problem caused by including u_0 can also be detected from the diagnostic trace plots. The trace plots from the model including u_0 shows a systematic pattern, which is a sign of lacking of convergence.

A similar problem happens to one of the predictor variable *race/ethnicity*, which takes the value of Asian, White, Hispanic, native American and other races, trace plots suggest the lack of convergence caused by identifiability of the parameter. We split race/ethnicity into four variables: *Asian*, *White*, *Hispanic* and *other races*, and keep the first three races but left out *other races* as a reference category to make the three race/ethnicity categories identifiable.

The variance of the ϵ_i (σ_ϵ^2) is identifiable because we fixed the bin boundaries. Our response data Y_i is partially ordered, the boundaries of each category γs is known, thus the variance σ_ϵ^2 is identifiable. If the boundaries $\gamma_1, \dots, \gamma_{J-1}$ are unknown, to ensure parameter

identifiability, the variance σ_ϵ^2 has to be fixed. For this reason, Albert and Chib (1993) imposed restriction on the bin boundaries by taking $\gamma_1 = 0$ and made $\sigma_\epsilon^2 = 1$. For partially ordered cases, a similar practice can be used. For a greater discussion of identifiability in latent probit models see Huang and Bandeen-Roche (2004), Xu and Craig (2009), and Qu et al. (1996).

4.9 Simulation Study

To assess statistical properties of this model in terms of estimation and prediction, we designed a simulation study. We consider the case of $k = 2$ fixed effects for $\boldsymbol{\beta}$ and $J = 12$ different random effects for $\boldsymbol{\tau}$. Specifically, we fixed $\boldsymbol{\beta} = (-1.45, 4.89)$, the random effect $\boldsymbol{\tau} = (81.20, 79.67, 81.80, 81.10, 75.68, 78.85, 77.14, 76.36, 76.12, 78.00, 80.91, 78.44)$, and standard deviation of error terms $\sigma_\epsilon = 6.08$. A vector for instructors was made with random assigned number 1-12 of probabilities for each instructor $c(0.0896, 0.0535, 0.0928, 0.2170,$ $0.0865, 0.0401, 0.0275, 0.1352, 0.1195, 0.0448, 0.0464, 0.0472)$. Based on this, the design matrix for instructors $\boldsymbol{Z}$ was set. Each created observation was randomly assigned an $\boldsymbol{X}_i$ vector and assigned a group for the random effect ($\boldsymbol{Z}_i$). The first element of $\boldsymbol{X}_i$ was generated following a Bernoulli trial with probability of success 0.283. The second element of $\boldsymbol{X}_i$ was drawn from a normal distribution.

Conditional on these parameters, fixed design vector, and random design vector, each observation was generated a latent course score following Equation 4.1 $S_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{\tau} + \epsilon_i$. This in turn implies the ordered letter grades Y_i (A+ to F). To incorporate the partial ordering, 5% of students were assigned to earn CR/NC. This process was repeated for sample sizes of $n = 500, 1000, 3000, 5000$ and $10,000$.

For each data set, we fit the model in Equation (4.1) with the prior distributions of Section 4.5.1 using the algorithm of Section 4.6. We obtained one hundred thousand post-convergence draws from the marginal distribution of each parameter and sample size. Fig-

ure 4.4 displays kernel density estimates of the marginal posterior distribution of β_1 and τ_7 for each sample size. The solid black vertical line represents the true value.

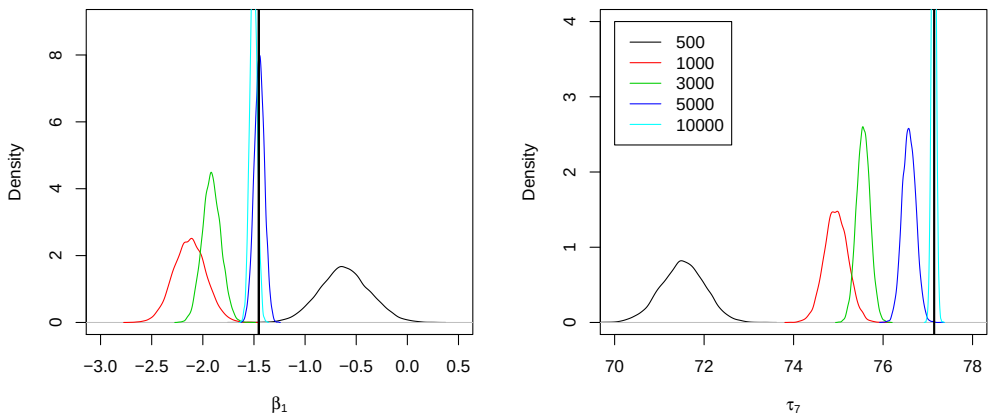

Figure 4.4: Density estimates of β_1 and τ_7. The left and right panel display the kernel density estimates of the marginal posterior distribution of β_1 and τ_7 respectively. Each color denotes a different sample size. The true value is represented by the solid black line.

In both panels of Figure 4.4, the posterior distributions are more variable at the smaller sample sizes. The posterior distribution from the smaller samples are also further from the true value, although they are unbiased estimates of the true value. As the sample size increases, the posteriors concentrate more around the true value. The plots of the other parameter estimates were so similar that we omitted them from the dissertation.

4.10 Application to our Redesigned Course Evaluation on Fall 2012 Data

We apply our partially ordered probit and logit models to the Redesigned course evaluation in Fall 2012. This section describes the parameter estimates of the models and the interpretation, as well as the prediction with these models.

4.10.1 Results

Returning to the course redesign evaluation at the University, we would like to apply the partially ordered multinomial probit and logit models to analyze the data. We include the variables in Table 4.13 based upon the results from stairwise elimination, in which the interaction terms do not suggest to be significant predictors, so the interaction terms are not included in the model. We implement the model stated in Section 4.5 and obtained 1,000,000 post convergence draws from the posterior distribution of the partially ordered probit model using the algorithm described in Section 4.6. Importance sampling was used to estimate the posterior mean of all parameters in the logit model. Table 4.13 displays the estimated posterior mean of all $\boldsymbol{\beta}$ parameters (fixed effects) for the probit and logit models. Figure 4.5 displays a distribution of the estimated marginal posterior of the $\boldsymbol{\beta}$ parameters in the probit model.

Variable	Probit	Logit
SAT/ACT	3.732	3.881
Course load	−0.618	−0.630
Semester GPA	10.213	10.493
HS grad. years	−0.239	−0.224
High school	3.878	3.146
Gender	0.303	0.349
Asian	5.260	6.507
White	3.272	3.486
Hispanic	1.795	2.115

Table 4.13: Estimated posterior means of the elements of $\boldsymbol{\beta}$. The first column contains the variable name associated with each β, and the column under "Probit" and "Logit" are the estimated posterior means of the elements of $\boldsymbol{\beta}$ for each independent variable, respectively. Estimated posterior means of the elements of $\boldsymbol{\beta}$.

From Table 4.13, we see that the parameter estimates from the two models have the same sign and are not far from each other in magnitude, we just use the probit model to illustrate the interpretation, since the interpretation of logit model is similar. The simple interpretation of these models is one of the benefits of including a latent variable into the model. The coefficients for *SAT/ACT* and *Semester GPA* are positive and large, which means students who did well in SAT or ACT math or have a higher GPA in the semester from other course tend to do better in Intermediate Algebra. Specifically, controlling for other variables, if students' standardized SAT/ACT math score increase one point, their Intermediate Algebra course score tend to increase 3.732 points.

Recall that students' *Semester GPA* ranges from 0 to 4.33, and was calculated on the courses excluding Intermediate Algebra. The parameter estimate of *Semester GPA* means that with controlling for other variables, if students' *Semester GPA* increases 1 unit, their Intermediate Algebra course score (ranges from 0 to 100) on average increase by 10.213 points. *Course load* has a negative coefficient, indicating that students who took more courses tend to do worse in Intermediate Algebra. If we hold other variables constant,

students taking one more credit hours will be likely to have 0.618 points lower in Intermediate Algebra, most of college courses have 3 credit hours, so if a student take 3 more credit hours, his or her Intermediate Algebra course score will decrease about 2 points ($0.618 \times 3 = 1.854$).

Gender has a coefficient of 0.303, since male is represented as 0 and female as 1, the coefficient 0.303 means that female students have 0.303 points higher than male students in Intermediate Algebra if other variables take the same values. It suggests that *Gender* is not a significant predictor variable.

HS grad. years has a negative estimated coefficient (-0.239), meaning that while controlling other variables, if a student graduated from high school one year earlier, his or her Intermediate Algebra course score will tend to be 0.239 points lower. Remember that the mean of *HS grad. years* for Intermediate Algebra students in Fall 2012 is 2.6 years, and *HS grad. years* ranges from a negative number to over 30 years. With the same other conditions, a student in Intermediate Algebra class who graduated from high school 5 years longer by the time taking the Intermediate Algebra course in August 20, 2012, can have an Intermediate Algebra course score 1 point lower, and 2 points lower if graduating from high school 10 years longer.

High school graduates on average had 3.878 points higher in Intermediate Algebra than GED certificate holders if we control for other variables. Asian students tend to do a little bit better than whites (2 points on average), whites are like to do better than Hispanics (about 1.5 points higher on average), and Hispanics do better in Intermediate Algebra than other races (about 1.8 points). The group *Asian* has a big variance, and this is caused by the small sample size: there were only 33 Asian students enrolled in Intermediate Algebra course in the Fall semester of 2012.

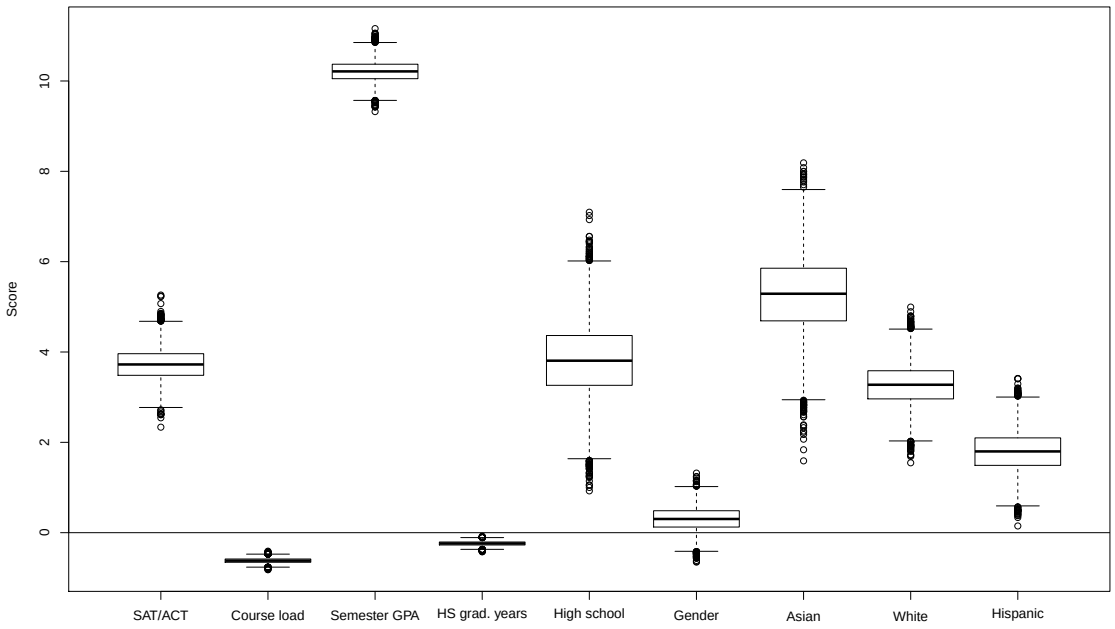

Figure 4.5: Boxplots of $\boldsymbol{\beta}$ estimated using draws from the posterior distribution of the probit model. The solid horizontal line is at zero.

Figure 4.6 displays boxplots of the marginal distribution of the instructor random effects under the probit model. The random effects were centered about zero for easier comparison. Recall that we were primarily interested in how students in the Redesigned course performed relative to other students, and the Redesigned course was taught solely by instructor 12, who did not teach any Traditional lecture sections in the semester. Thus, the difference of instructor random effects between instructor 12 and other instructors represents the difference of students' performance between Traditional lecture sections and the Redesigned course. Looking at the box plots, we see that the model estimates of the grade distribution in the Redesigned course (listed as Instructor 12) is similar to the instructors with the highest observed grade distributions (instructors 9 and 5), and are higher than that of most of the other instructors. The average score of the Redesigned course was about 8 points higher than the average student score of instructor 2. Students from the Redesigned course averaged approximately 4 points higher in the course (out of 100) than an average instructor (controlling for all of the covariates), which is almost half of a letter grade.

Using this model, we estimate that students in the Redesigned course averaged almost half a letter grade higher in Intermediate Algebra than students in a typical lecture section. This suggests that students in the Redesigned course did as well as students from the sections of the best scored instructor with a Traditional lecture method.

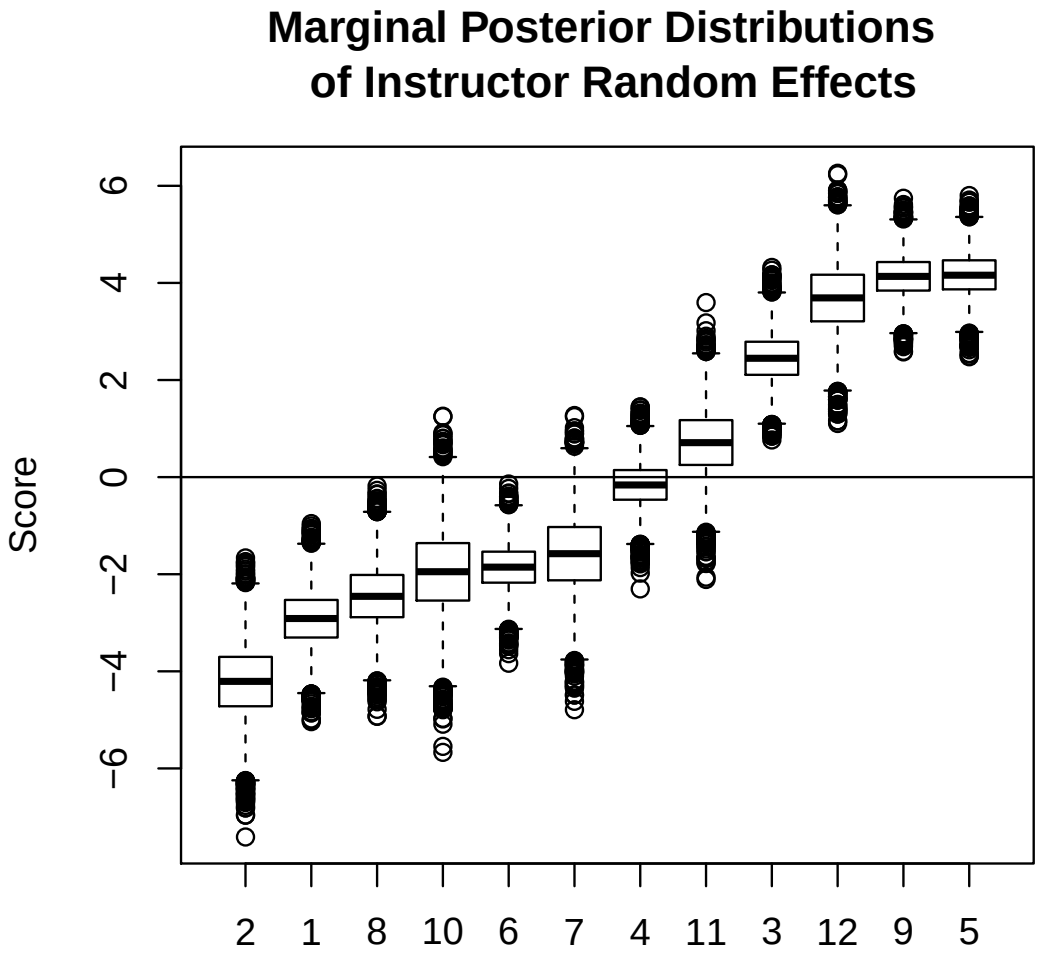

Figure 4.6: Boxplots of instructor random effects ($\boldsymbol{\tau}$). The draws were centered at zero by subtracting the overall mean (of all instructor effects). Instructor 12 is the Redesigned course and the numbers 1-11 represent different instructors who taught the Traditional lecture sections.

4.10.2 Prediction with The Partially Ordered Probit and Logit Models

An advantage of fitting models through Markov Chain Monte Carlo approaches is that we can easily make predictions about the course score for individuals. Considering a student who is a white, male, high school graduate with average SAT/ACT score (ACT math score 19 or SAT mathematics score 463), average $Semester\ GPA$ (2.97), average number of years since high school graduation (2.6 years) and who was in the Redesigned course. Using our fitted probit model, his average predicted score is 80.3 and his probability of passing the Intermediate Algebra course is 0.75. Figure 4.7 displays the posterior predicted grade distribution for this student under both the probit and logit model.

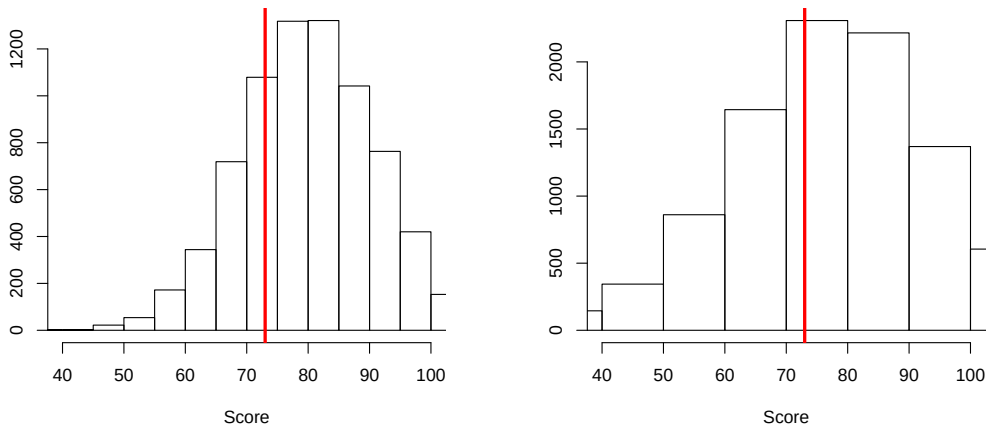

Figure 4.7: Estimated posterior predictive distribution of scores for an example student from the Redesigned course section under the probit (left figure) and logit (right figure) model. The vertical red line represents the passing score 73.

If this student was instead taught by instructor 2 (the instructor with the lowest students' score), his average posterior predicted score under the probit model would change to 72.2 while his probability of passing the Intermediate Algebra course drops to 0.47. Figure 4.8 displays the posterior predicted grade distribution for this student under both the probit and logit model.

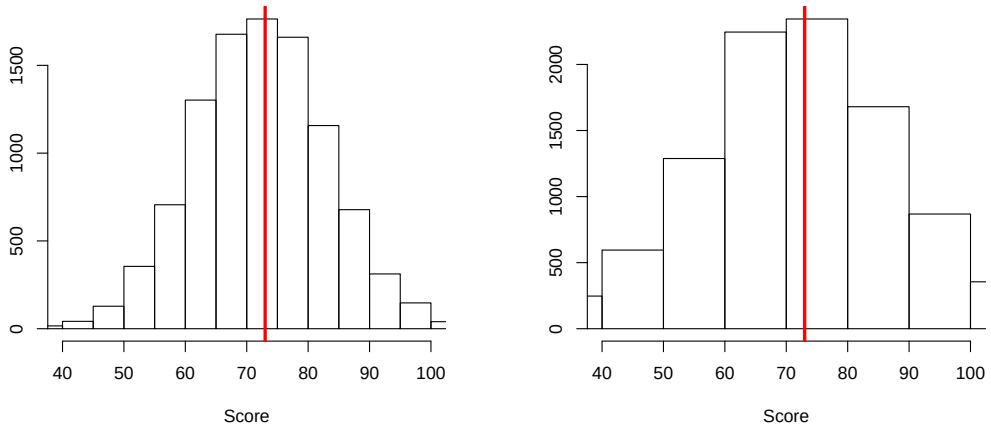

Figure 4.8: Estimated posterior predictive distribution of scores for an example student from a Traditional lecture section: the section taught by instructor 2 under the probit (left figure) and logit (right figure) model. The vertical red line represents the passing score.

If the model assumptions are reasonable, the ability to quickly make predictions of student scores allows us to quantify how much better a section would perform if the format was changed to the Redesign.

4.11 Comparison to Alternative Analysis

The main advantage of this model is that we can use all of the data while a standard ordered or binary probit model cannot. In our example, the binary probit model ignores the magnitude of the letter grade by converting them to passing or failing grades and completely ignores those who receive a W, I or AUD. The ordered probit model similarly ignores grades of CR, NC, W, I, AUD. The ordered probit uses the least amount of observations but has more detailed information than the binary probit model. The partially ordered probit model uses all of the observations with all of the detail.

We divide the data into five equal subsets, each time using one subset as training set with which we estimate parameters of the model, while leaving the other four subsets as test set. We plug in the value of the covariates of each observation to the model we obtained from the training set, and predict the value of the response, then compare with the observed response value. Cross validation can test how well the model will fit an independent data set, and provide unbiased assessment of the model estimate. The mean squared error (MSE) is one of many ways to quantify the difference between values predicted by an estimator and the true values of the response.

In signal detection theory, a Receiver Operating Characteristic (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system. It is created by plotting the fraction of true positives out of the total actual positives (true positive rate) versus the fraction of false positives out of the total actual negatives (false positive rate), at various threshold settings. The Area Under ROC Curve (AUC) analysis provides tools to select possibly optimal models and to discard suboptimal ones independently.

To quantify the impact of this loss of information, we utilized five-fold cross validation with two measures of model fit: a mean squared error (MSE) and the area under the ROC curve (AUC). These measures are based on only predicting the probability of passing the course. We chose this because the University is most interested in predicting pass rates. For the ordered probit and binary probit, we used the Bayesian formulation from Albert and Chib (1993) with the same random effects and predictors as the partially ordered model.

We applied a Bayesian version of the partially ordered probit model, ordered probit model and binary probit model on the Fall 2012 Intermediate Algebra student data. The data was randomly placed into 5 folds of equal size. For each of the five folds, the data not in that fold was used to fit each of the three probit models. Predictions of the passing rate were made for each student in that fold. From those predictions, the MSE and AUC were calculated with the actual pass/fail results for those students. Students who received a W, I, or AUD were excluded at this stage. Table 4.14 presents the average MSE and AUC over the five folds for all three models.

Model	MSE	AUC
Partially Ordered Probit	0.310	0.654
Ordered Probit	0.327	0.599
Binary Probit	0.348	0.642

Table 4.14: Estimated measures of model fit for each model. The first column lists the models used. The second column gives an estimated MSE using five-fold cross validation. The third column gives an estimated AUC using five-fold cross validation.

We see from Table 4.14 that, in terms of MSE and AUC, the partially ordered model performs the best. In other words, the information gained from using all grading information for students does translate into a better predictive model. The ordered model performs worse than the binary probit model for this data in terms of AUC but better in terms of MSE. We attribute this to it being unable to use the 158 students who received a CR or NC. The extra detail of the ordered model could not compensate, in terms of model fit, for the loss of these students. If there were no CR/NC students, we suspect that the ordered model would

perform better than the binary probit model.

Chapter 5

Analysis of Spring 2013 Course Redesign Data

5.1 Introduction

In the Spring semester of 2013, all the Intermediate Algebra students at the University were put into the Redesigned course. We apply our partially ordered probit and logit models, presented in Chapter 4, to the Intermediate Algebra Redesigned course grade data of Spring 2013. We use the same variables with the Spring 2013 data as those with the Fall 2012 data. After cleaning the data, we find that there were 950 students enrolled in Intermediate Algebra in the Spring semester of 2013, and we use the information of these students for analysis.

5.2 Descriptive Statistics

5.2.1 Descriptive Statistics for the Spring 2013 data

We would first like to explore the data with tables and Chi-Square tests to answer the following questions that we are interested in about the Spring 2013 Intermediate Algebra student success.

Distribution of letter grades in the Redesign of Spring 2013

Since 54.63% of students received Incomplete in the Redesign in the previous semester, which tremendously influenced our method of analysis, we would like to know the grade distribution of Spring 2013 Intermediate Algebra students in the Redesigned course. We see from Table 5.1 that 56.21% (534 out of 950) of the students received an "I", compared to 54.63% (118 out of 216) in the Redesign of the previous Fall semester. 15.05% students got a "W". Oddly, 3 students preferred to get D or D- instead of a "W" or "I", although the Department of Math and Statistics planned not to give D or F grades in the Redesign. There were 2 students registered for "CR/NC", and both ended up with a "NC". Because there were so many Incompletes, we must take this into account when building models.

Pass Rate of students with and without SAT/ACT scores on the Spring 2013 data after removing "I" and "AUD"

Students with an "I" have not completed the course for this analysis, and there were 56.21% (534) students received a "I" in the Spring 2013, so we only considered students who have finished or withdrawn from the course. We exclude the "I" and "AUD" students from the dataset at this stage only for comparing pass rate. Later for fitting the multilevel logistic model and the partially ordered probit/logit models, we will use all of the data including

Letter grade	Frequency	Percent
A+	18	1.89
A	77	8.11
A-	36	3.79
B+	26	2.74
B	71	7.47
B-	16	1.68
C+	7	0.74
C	11	1.16
D	1	0.11
D-	2	0.21
CR	5	0.53
NC	2	0.21
W	143	15.05
I	534	56.21
AUD	1	0.11

Table 5.1: Distribution of letter grades in the Redesign of Spring 2013, the first column represents each letter grade was given in the Redesign in Spring 2013; the second column gives the the number of students who received each letter grade in the Redesign of spring 2013, and the third column provides the precentage of students who received each letter grade listed in the first column.

the "I" and "AUD" students. Students with "W" grade stay in the data in this step because "W" is considered as a fail in the Redesign. After deleting students with "I" and "AUD" from the dataset, we have 415 observations. We treat the letter grades: A, B(with + and -) and $C+, C$ and CR as passing grades, and W as failing grade. Table 5.2 tells that there were 267 out of 415 (64.34%) students passed, and the pass rate for students with an SAT/ACT score is 66.40%; the pass rate for students without an SAT/ACT score is 47.83%. Among the students who took the SAT/ACT, 33.6 % of them did not pass, and 52.17% students who do not have a SAT/ACT score failed the course.

The p-value for the Chi-Square test with degree of freedom 1 is 0.0132, suggesting that the pass rate of the students with an SAT/ACT score is significantly higher than those without an SAT/ACT score. Again, this indicates that students who transferred from other

	Students with SAT/ACT	Students without SAT/ACT	Total
Passed students	245	22	267
Passed percentage among all students	59.04	5.30	64.34%
Pass rate among each group	**66.40**	**47.83**	
Failed students	124	24	148
Failed percentage among all students	29.88	5.78	35.66%
Failing rate among each group	33.60	52.17	
Total number of students in each group	369	46	415
Percentage of students in each group over all	89.92	11.08	100%

Table 5.2: Pass rate comparision between students who took an SAT/ACT and those who did not. The "Passed percentage among all students" means the percentage of passed students in each group over all students enrolled in the Intermediate Algebra course in the Spring semester of 2013, similarly, the "Failed percentage among all students" means the percentage of failed students in each group over all students enrolled in the Intermediate Algebra course in the Spring semester of 2013. The Pass (Failing) rate represents the percentage of passed (failed) students in each group (students with an SAT/ACT score or those without).

universities did not do as well as students who came to the University directly from high school.

Pass rate of students across different race/ethnicity

The University has a large, diverse student population. In particular, the University has a much larger Hispanic population in comparison to the rest of the country. In the Intermediate Algebra course, Hispanic has the largest student population, and always over 50% of the students in this course are Hispanic; only a very small percentage of students are from the race/ethnicity groups other than Hispanic and whites. We would like to continue to

monitor students' success across different race/ethnicity in the Intermediate Algebra course in Spring 2013, while the new teaching method was fully adopted, i.e. completely using computer software to teach Intermediate Algebra.

The pass rates of each race/ethnicity group in the Redesign in the Spring semester of 2013: 64.14% for Hispanic, 64.58% for whites, and 61.29% for native Americans, 64.29% for two or more races. There is only one native Hawaiian student in the data, and there are 4 non-resident aliens, 2 Asian students, 8 students with "race/ethnicity unknown", 8 black or African Americans. The sample size for these races are too small, so the pass rates yielded from these race/ethnicity groups are not reliable. The number of passed or failed students from these race/ethnicity groups are less than 5, which causes inaccuracy of Chi-square test, so we deleted them from the data for the sake of Chi-Square test. To protect the privacy of students from these very small race/ethnicity groups such as non-resident aliens, Asian, black or African Americans and native Hawaiian, the table of pass rate across the race/ethnicity groups is omitted.

The p-value for the Chi-square test on the data without the very small race/ethnicity groups is 0.9899, suggests that there is not significant difference in the pass rate among different race/ethnicity groups in the Redesign in Spring 2013. Recall that there was a significant difference in students' pass rate across different race/ethnicity groups in Fall 2012.

According to the University "Official Enrollment Report Spring 2013" (UNM, 2013), there were 40.92% of whites, 38.33% of Hispanic students, and small percentages of other races. Thus, there was a bigger percentage of Hispanic students in Intermediate Algebra than in the University at large.

Student success between high school graduates and GED certificate holders

Again, we would like to investigate how the students with GED certificate did in the Intermediate Algebra course compared to high school graduates, since most of GED certificate holders are composed of home-schooled students, or students who did not finish high school but enrolled in college later. Table 5.3 displays student success between high school graduates and GED certificate holders in the Redesign in Spring 2013 on the data excluding "I" and "AUD" students. Among high school graduates, 65.39% students passed, and in the group of GED certificate holders, 45.45% students passed. There were only 22 GED students in the data, while 393 high school graduates. The p-value for Chi-Square test is 0.0574 with 1 degree of freedom, suggesting that the pass rate is not significantly different at 0.05 significance level between students who graduated from high school and those who have GED certificate. We also did the Chi-Square test on the data with all Spring 2013 students including "I" and "AUD" students, the results were same: the pass rates of high school students and GED holders are not significantly different.

Pass Rate between males and females

Our analysis on the Fall 2012 Intermediate Algebra student data concluded that female students did better than male students in this course. We would like to redo the test using the Spring 2013 data. Table 5.4 provides the pass rate between male and female students. There were 66.40% of female students passed Intermediate Algebra in Spring 2013, and 61.11% males passed the course. Again, the p-value for Chi-Square test in pass rate between male and female students is 0.2722, suggesting that the pass rate between male and female is not statistically different. Note that there were 578 (60.84%) females, and 372 (39.16%) males enrolled in the Intermediate Algebra class at the University in Spring 2013, and this is consistent to the student population at the University, since there were more female than male enrolled in the University in Spring 2013. According to the University "Official

	High school graduates	GED certificate holders	Total
Passed students	257	10	267
Passed percentage among all students	61.93	2.41	64.34%
Pass rate among each group	**65.39**	**45.45**	
Failed students	136	12	148
Failed percentage among all students	32.77	2.89	35.66%
Failing rate among each group	36.41	54.55	
Total number of students in each group	393	22	415
Percentage of students in each group over all	94.70	5.30	100.00

Table 5.3: Passe rate between high school graduates and GED certificate holders in the Redesign in Spring 2013. The "Passed percentage among all students" means the percentage of passed students in each group over all students enrolled in the Intermediate Algebra course in the Spring semester of 2013; similarly, the "Failed percentage among all students" means the percentage of failed students in each group over all students enrolled in the Intermediate Algebra course in the Spring semester of 2013. The "Pass (Failing) rate" represents the percentage of passed (failed) students in each group (high school graduates or GED certificate holders) enrolled in the course in Spring 2013 over all students in each group.

Enrollment Report Spring 2013" (UNM, 2013), there were 19,464 undergraduate students enrolled in the Spring, 2013, consisting of 8,607 males and 10,857 females.

Distribution of quantitative independent variables

Figure 5.1 describes the distribution of independent variables: $HS\ grad.\ years$ (high school graduation years), SAT/ACT (standardized SAT/ACT score), $Course\ load$ and $Semester\ GPA$. We see that high school graduation years is strongly right skewed, and about 95% of students graduated from high school less than 15 years ago. The mean high school graduation years is 3.11 years, while the average high school graduation years of

	Female	Male	Total
Passed students	168	99	267
Passed percentage among all students	40.48	23.86	64.34%
Pass rate among each group	**66.40**	**61.11**	
Failed students	85	63	148
Failed percentage among all students	20.48	15.18	35.66%
Failing rate among each group	33.60	38.89	
Total number of students in each group	253	162	415
Percentage of students in each group over all	60.96	39.04	100.00

Table 5.4: Passe rate between male and female students in the Redesign in Spring 2013. The "Passed percentage among all students" means the percentage of passed students in each gender over all students enrolled in the Intermediate Algebra course in the Spring semester of 2013; similarly, the "Failed percentage among all students" means the percentage of failed students of each gender over all students enrolled in the Intermediate Algebra course in the Spring semester of 2013. The "Pass (Failing) rate" represents the percentage of passed (failed) students of each gender enrolled in the course in Spring 2013 over all male or female students who did not receive an I and a AUD in the course in Spring 2013.

students in Intermediate Algebra course in Fall 2012 was 2.602 years, standard deviation is 5.84. The mean SAT/ACT is -0.705, which is lower than that of students in the Fall (-0.458 with standard deviation 0.4239), we don't know if that means students are doing worse in math in high school, or the ACT/SAT math tests becomes harder, or many students who had higher SAT/ACT score but still in the range of SAT 450-500 or ACT 19-21 took Intermediate Algebra in the local community college for a lower cost, and standard deviation is 0.406. The histogram for *Course load* tells that the average credit hours each student took in Spring 2013 was 13.069 credit hours, comparing to 13.97 credit hours in the Fall. Note that these credit hours don not include the three credit hours from Intermediate Algebra. The distribution of *Semester GPA* is roughly normal with mean 2.7085, compared to average

semester GPA 2.97 in the Fall semester of 2012.

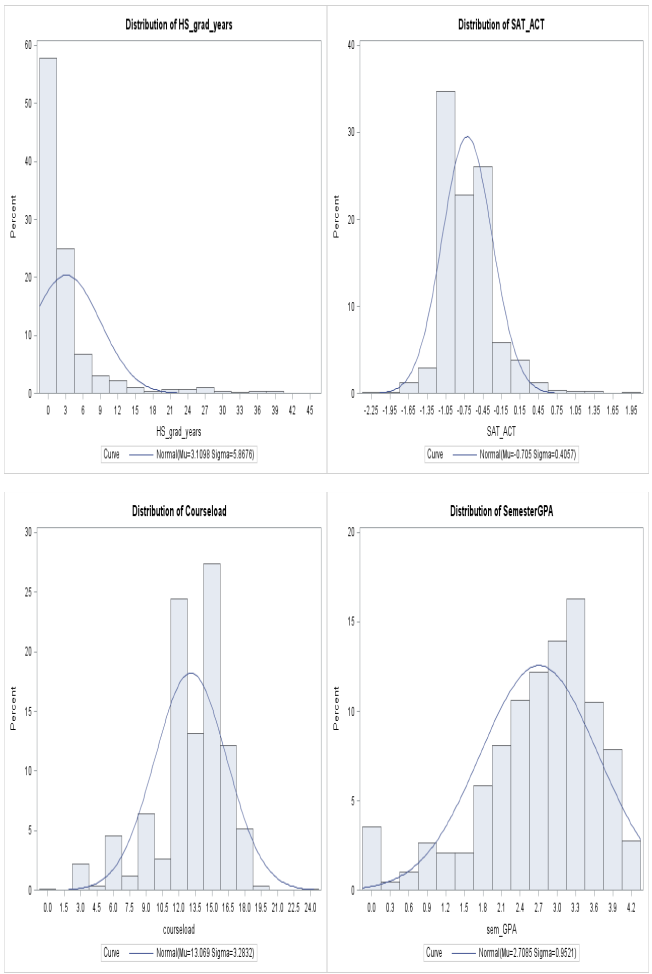

Figure 5.1: Histograms for quantitative independent variables about Intermediate Algebra students from Spring 2013. The left upper panel displays the distribution of the elapsed time in years after students graduated from high school by taking Intermediate Algebra in Spring 2013. The right upper panel: the distribution of the SAT/ACT score of the students enrolled in the course from Spring 2013. The left lower panel presents the distribution of the total credit hours excluding the credit hours from Intermediate Algebra. The right lower panel gives the distribution of students semester GPA calculated excluding the three credits from Intermediate Algebra in Spring 2013 when they took Intermediate Algebra.

Students enrolled in both of the semesters

All Intermediate Algebra students were in the Redesign in Spring 2013, while there was only a pilot section of 216 students of the Redesign, and most of other students were taught with the Traditional lecture method in Fall 2012. In order to compare student learning performance in the Redesign between Fall 2012 and Spring 2013, as well as comparing student learning performance between Spring 2013 and that of the Traditional lecture sections in Fall 2012, we need to combine the datasets of the two semesters together.

After getting rid of the 12 students who did not receive any grade, which is probably a recording mistake, there are 2258 students in the dataset. Among them, 166 students were recorded twice because they were enrolled in the Fall and retook the course in the Spring. From these 166 duplicates who both enrolled in the Fall and the Spring, we pulled out students who were in the Redesign of Fall 2012, and found that there were 11 students who took Intermediate Algebra in the Redesign of Fall 2012. In other words, there are 11 students took Intermediate Algebra in the Redesign in Fall 2012, and re-enrolled the course in Spring 2013, and 155 students taught by the Traditional lecture method in Fall 2012 retook it in the following Spring semester of 2013. Note that the students who received Incomplete do not have to register the course again, they just need to continue to work on the materials they left in the previous semester, and after they finish and pass the course, the instructor will go back to their Fall 2012 grade and change the "I" to a passing grade. If the Incomplete students can not pass the course in three semesters (including summer), the "I" will become an "F".

Table 5.5 describes the Fall grades distribution of these 11 students who did not pass the course in the Redesign in Fall 2012 and continued to work on it in Spring 2013. We do not know why it shows that the two Incomplete students re-enrolled in the course. Recall there were 118 (54.63%) Incompletes in the Redesign and 8 Incompletes from the lecture sections, so 126 students total received an "I" in Fall 2012. These 126 students should be continuing

to work on the course in Spring 2013. We do not have information about their score or their updated grade to change the "I" by the end of Spring 2013.

Letter Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
W	7	63.64	7	63.64
D	2	18.18	9	81.82
I	2	18.18	11	100.00

Table 5.5: Fall grade distribution of students retaking Intermediate Algebra in the Spring that were in the Redesign in Fall 2012.

There were a total of 564 students who did not pass in Fall 2012 (414 from the lecture sections, and 150 students from the Redesign), only 166 of them continued in the Spring. Again, we took a subset from 166 duplicates, and only kept students' Spring grade. Table 5.6 gives the grade distribution of the Spring grade for the 166 students who did not pass in the Fall and continued in the Spring. From our past research, whether or not students attempted Intermediate Algebra before is not significant in predicting their learning performance in the current semester, and students who tried to take Intermediate Algebra before do slightly worse on average than students who never tried to take the course before. The trend still holds on the Spring data, as we can see from Table 5.6, 66.87% of re-enrolled students received "I", the pass rate for these re-enrolled students is as low as 10.26% if we consider "I" as a fail for the current semester.

To compare the student performance in Spring 2013, the pilot Redesign in Fall 2012, and the Traditional lecture sections in Fall 2012, considering the pass rate in the Intermediate Algebra course was consistently hovering below 50%, so there were always returning students in this course. There were 166 students in the Spring 2013 who returned after they failed the course in Fall 2012, while there were definitely also students who failed the course before the Fall semester of 2012, and returned from the previous semesters, but we do not have information of Intermediate Algebra students before Fall 2012. Because the Intermediate Algebra grades of these three groups of students are all under the same influence of some

Letter Grade	Frequency	Percentage
A	6	3.61
A-	3	1.81
B+	1	0.60
B	4	2.41
B-	1	0.60
C	2	1.20
I	111	66.87
W	38	22.89

Table 5.6: Spring grades of the 166 students who did not pass in the Fall but continued in the Spring. There were 89.76% of this group of students did not pass by the end of Spring 2013

students repeatedly taking the course and the information we have, we could only ignore this effect for right now.

5.2.2 Descriptive statistics on combined data

Before we fit the models to the data, we would like to explore the combined data for the Fall and the Spring, to see the effect of the predictor variables on a larger population to get a more accurate picture. In other words, how students' success in Intermediate Algebra differ between the students who have an SAT/ACT score and those who do not, high school graduates and GED holders, as well as different gender and race/ethnicity groups. In this section, we still treat the grades A, B(with +/-), $C+$, C and CR as passing grade, and all other grades including $C-, D+, D, D-, F, NC, W, I, AUD$ as failing grades.

Difference in Intermediate Algebra pass rate between students with an SAT/ACT score and those without

Table 5.7 provides the pass rate of students in Intermediate Algebra between students with an SAT/ACT score and those without. The pass rate of students with an SAT/ACT score in this course is 46.30%, while the pass rate of students without an SAT/ACT score is 30.59%. There were only a small percentage of students (9.70%) who did not take the the SAT/ACT. The p-value of Chi-Square test <.0001 suggests that the pass rate of students with an SAT/ACT score in this course is significantly higher than those without an SAT/ACT score. Using the students from the two semesters for analysis, we still get the same conclusion.

Gender effect on student Intermediate Algebra success

From the combined data of Intermediate Algebra students from the two semesters, Table 5.8 tells that there are more females than males (57.17 % vs. 42.83%) enrolled in the Intermediate Algebra course, and there were also more females (46.01%) than males (43.13%) passed the course. Again, the difference in the pass rate between females and males in the course is not significant since the p-value from the Chi-Square test is 0.1721, bigger than 0.05 significance level. This suggests that students' math ability is not significantly different between males and females.

High school effect on student Intermediate Algebra success

Table 5.9 presents the student Intermediate Algebra success between high school graduates and GED certificate holders. We see that 95.57 % students enrolled in Intermediate Algebra are high school graduates, only 4.43% (100 students) have GED certificate. The pass rate for high school graduates in Intermediate Algebra course of the two semesters is 45.60%, and 27.00% for GED certificate holders. The p-value of Chi-Square test is 0.0003, which suggests

	Students with SAT/ACT score	Students without SAT/ACT score	Total
Passed students	944	67	1011
Passed percentage among all students	44.81	2.97	44.77%
Pass rate among each group	46.30	30.59	100%
Failed students	1095	152	1247
Failed percentage among all students	48.49	6.73	55.23%
Failing rate in each group	53.70	69.41	
Total number of students in each group	2039	219	2258
Percentage of students in each group over all	90.30	9.70	100%

Table 5.7: Pass rate comparision between students took an SAT/ACT and those who did not. The "Passed percentage among all students" means the percentage of passed students in each group over all students enrolled in the Intermediate Algebra course in the two semesters. Similarly, the "Failed percentage among all students" means the percentage of failed students in each group over all students enrolled in the Intermediate Algebra course in the two semesters. The "Pass (Failing) rate" represents the percentage of passed (failed) students in each group (students with an SAT/ACT score or those without in the course) over all students in each group.

that this difference is statistically significant.

Pass rate between different race/ethnicity groups

Recall that we had to take away some race/ethnicity group to use the Chi-square test because of very few students in these race/ethnicity groups when we analyzed the Fall 2012 or Spring 2013 data. Redoing the test on more students from the two semesters will give use a more reliable result. Table 5.10 describes the pass rate between different race/ethnicity. The pass rates of Asian and non-resident alien are the biggest, 55.81% and 53.33%, respectively, the

	Female	Male	Total
Passed students	494	417	1011
Passed percentage among all students	26.31	18.47	44.77%
Pass rate for each group	**46.01**	**43.13**	
Failed students	697	550	1247
Failed percentage among all students	30.87	24.36	55.23%
Failing rate in each group	53.99	56.88	
Total number of students in each group	1291	967	2258
Percentage of students in each group over all	57.17	42.83	100.00

Table 5.8: Passe rate between female and male students in the two semesters. The "Passed percentage among all students" means the percentage of passed students of each gender over all students enrolled in the Intermediate Algebra course in the two semesters. Similarly, the "Failed percentage among all students" means the percentage of failed students of each gender over all students enrolled in the Intermediate Algebra course in the two semesters. The "Pass (Failing) rate" represents the percentage of passed (failed) female (or male) students over all female (or male) students.

next is whites (51.22%), two or more races (44.19%), Hispanic (43.19%), native American (36.18%), black or African American (36.14%), then "Race unknown" (31.25%). Because the percentages of students from two or more races, black or African American, Asian, "Race unknown", and non-resident alien are so small among the Intermediate Algebra student population, range from 3.81% to 0.67%, we put these minority races in the group of "other races" to fit the models. The p-value from Chi-Square test is 0.0014, suggests that there is a significant difference in Intermediate Algebra pass rate between different race/ethnicity groups.

	High school graduates	GED certificate holders	Total
Passed students	984	27	1011
Passed percentage among all students	43.58	1.20	44.77%
Pass rate in each group	**45.60**	**27.00**	
Failed students	1174	73	1247
Failed percentage among all students	51.99	3.23	55.23%
Failing rate in each group	54.40	73.00	
Total number of students in each group	2158	100	2258
Percentage of students in each group over all	95.57	4.43	100.00

Table 5.9: Passe rate comparison between high school graduates and GED certificate holders from the two semesters. The "Passed percentage among all students" means the percentage of passed students in each group over all students enrolled in the Intermediate Algebra course in two semesters. Similarly, the "Failed percentage among all students" means the percentage of failed students in each group over all students enrolled in the Intermediate Algebra course in the two semesters. The "Pass (Failing) rate" represents the percentage of passed (failed) students in each group (high school graduates or GED certificate holders) enrolled in the course in the two semesters over all students in each group.

Distribution of quantitative independent variables on the combined Fall and Spring Intermediate Algebra grade data

Comparing the histogram of *HS grad. year* in Table 5.2 and Table 5.1, there were some high school students taking Intermediate Algebra in the Fall 2012, but not in the Spring 2013. The distributions of these variables are similar between the Fall data and the combined data of two semesters.

	Hisp.	White	Native Ameri.	Two or more races	Black or African Ameri.	Asian	race unknown	Non -Res Alien	Total
Pass	530	316	55	38	30	24	10	8	1011
1	23.50	14.01	2.44	1.69	1.33	1.06	0.44	0.35	44.83
2	43.19	51.22	36.18	44.19	36.14	55.81	31.25	53.33	
Fail	697	301	97	48	53	19	22	7	1244
3	30.91	13.35	4.30	2.13	2.35	0.84	0.98	0.31	55.17
4	56.81	48.78	63.82	55.81	63.86	44.19	68.75	46.67	
Total	1227	617	152	86	83	43	32	15	2255
5	54.41	27.36	6.74	3.81	3.68	1.91	1.42	0.67	100.0

Table 5.10: Pass rate of students from two semesters between different races. The row "Pass" gives the number of passed students from each race. The row starting with "1" provides the percentage of passed students in each race over all Intermediate Algebra students from the two semesters, row "2" describes the pass rate in each race based on the data from the two semesters. The row "Fail" gives the number of failed students, row "3" tells the percentage of failed students in each race over all Intermediate Algebra students from the two semesters. The row starting with "4" represents the failing rate in the course in each race for the students enrolled in the two semesters. "Total" means the total number of students in each race, and the row beginning with "5" gives the percentage of students of each race among all students enrolled in the course in the two semesters.

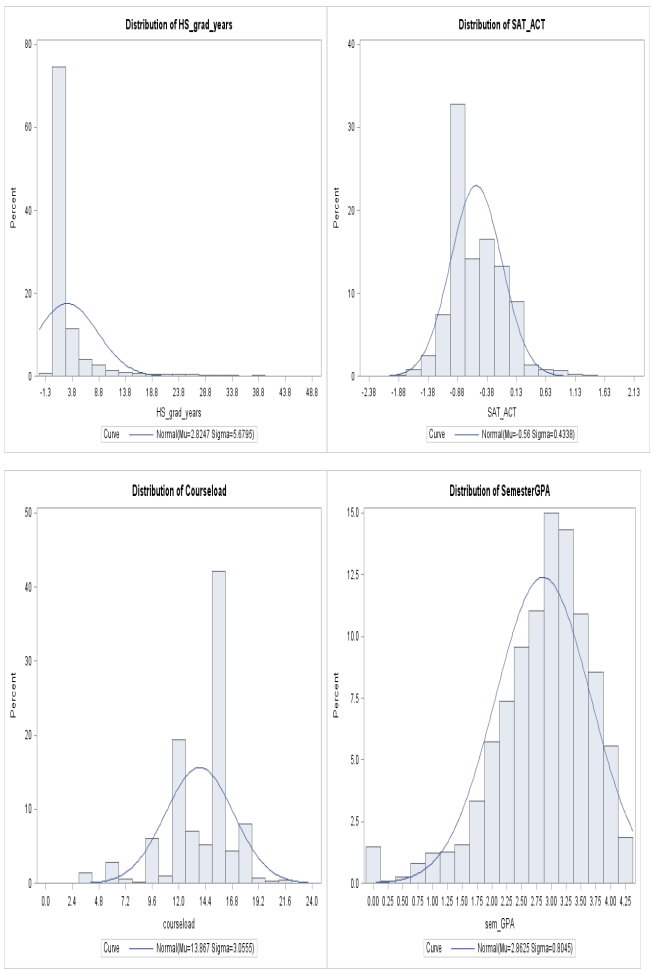

Figure 5.2: Histograms for the quantitative independent variables for Intermediate Algebra students from two semesters. The left upper panel displays the distribution of the elapsed time in years after students graduated from high school until taking Intermediate Algebra in Fall 2012 or Spring 2013; the right upper panel: the distribution of SAT/ACT score of the students enrolled in the course from the two semesters. The left lower panel presents the distribution of the total credit hours excluding the credit hours from Intermediate Algebra. The right lower panel gives the distribution of students' semester GPA calculated excluding the three credits from Intermediate Algebra in the Fall semester of 2012 or Spring 2013 when they took Intermediate Algebra.

Investigation on high school graduation years

From Figure 5.1 and Figure 5.2, we can only see that there was a very small percentage of students who graduated from high school more than 15 years ago. Because mathematics materials across courses are closely related, the higher level math builds upon the lower level, it is common that the longer students wait to take Intermediate Algebra, they forget more of what they have learned in high school, and consequently do worse in this course. In response to the feedback from some professors that very old students tend to do better than young people in school, it is necessary to do further investigation on the influence of the amount of time since students graduated from high school on the students learning outcomes on a much larger dataset.

Again, we split up *HS grad. years* (high school graduation years) to several categorical variables: students who are still in high school so have years graduated from high school less than 0; new high school graduates having years graduated from high school between 0 to 5; students who have graduated from high school longer than 5 to 10 years; students who graduated from high school more than 10 to 20 years ago; older people who graduated from high school more than 20 years to 30 years, 30 years to 40 years as well as longer than 40 years.

Table 5.11 displays that Intermediate Algebra pass rates of high school students and those who just graduated from high school are close (46.15% to 46.92%), but there is a dramatic drop in pass rate for the students graduate from high school 5 years to 10 years ago (44.92% to 33.33%), then this decreasing pace is a little bit slower for students graduated from high school 10 to 20 years ago (33.33% to 28.40%), and the decreasing trend tend to be steady after students graduated from high school for 10-20 years, 20-30 years, 30-40 years (28.40% to 27.5% to 25.0%); for the people graduated from high school more than 40 years, their pass rate increase to 33.33. However, there were only 9 students (0.40%) in this group. The p-value for Chi-Square test is less than 0.0001, suggest that students' pass rate does

differ with the increasing of high school graduation years. Since there were only 9 students who graduated from high school more than 40 years ago, we deleted these outliers from the dataset for fitting the models.

Years graduated from high school	<0	0-5	5-10	10-20	20-30	30-40	⩾40	Total
Passed students	12	907	51	23	11	4	3	1011
Passed percentage among all students	0.53	40.17	2.26	1.02	0.49	0.18	0.13	44.77
Pass rate in each category	**46.15**	**46.92**	**33.33**	**28.40**	**27.5**	**25.0**	**33.33**	
Failed students	14	1026	102	58	29	12	6	1247
Failed percentage among all students	0.62	45.44	4.52	2.57	2.08	0.53	0.27	55.23
Failing rate in each category	53.85	53.08	66.67	71.60	72.31	75.0	66.67	
Total number of students in each group	26	1933	153	81	40	16	9	2258
Percentage of students in each group over all	1.15	85.61	6.78	3.59	1.77	0.71	0.40	100.0

Table 5.11: Pass rate of students across students with different years graduated from high school. The "Passed percentage among all students" means the percentage of passed students in each group over all students enrolled in the Intermediate Algebra course in two semesters; similarly, the "Failed percentage among all students" means the percentage of failed students in each group over all students enrolled in the Intermediate Algebra course in the two semesters. The "Pass (Failing) rate" represents the percentage of passed (failed) students in each age group enrolled in the course in the two semesters over all students in each group.

5.2.3 The test for multicollinearity among predictor variables

We re-investigate the problem of multicollinearity among predictor variables on the combined data. See Section 4.2.3 for a discussion about multicollinearity. We generate a pairwise correlation matrix between all continuous variables, and did not see any high correlation between any two variables. Multicollinearity diagnostics tolerance, Variance Inflation Factor (VIF) and condition index were also assessed. The VIFs for whites and Hispanic are bigger than 2.5, making us a little concerned of multicollinearity among the predictor variables. Further, we checked the condition index. One condition index is quite large: 32.99447, pair this with a very small Eigenvalue (0.00612), and two big values for proportion of variation (0.97480 and 0.64010), these are evidence that there is the existence of multicollinearity. This is caused by the reference group, which includes the students other than whites and Hispanic, counts for a very small percentage (only about 10%). A solution to this multicollinearity is to increase the

Variable	Parameter Estimate	Standard Error	t Value	$Pr > \|t\|$	Tolerance	Variance Inflation
Semester GPA	0.12529	0.01317	9.52	$< .0001$	0.93770	1.06644
Course load	-0.00286	0.00405	-0.70	0.4812	0.89839	1.11311
SAT/ACT	0.15989	0.02542	6.29	$< .0001$	0.85183	1.17395
HS grad. years	-0.02567	0.00467	-5.50	$< .0001$	0.89140	1.12183
High school	0.00228	0.07951	0.03	0.9771	0.98350	1.01678
redesign	-0.13102	0.01153	-11.36	$< .0001$	0.86008	1.16269
White	0.05396	0.03684	1.46	0.1432	0.39582	2.52643
Hispanic	0.02250	0.03380	0.67	0.5056	0.37040	2.69982
native American	-0.00555	0.05102	-0.11	0.9134	0.67270	1.48654

Table 5.12: Multicollinearity dianostics. "redesign" is a variable telling wether a student took Intermediate Algebra in the Redesign in Fall 2012, Traditional lecture in Fall 2012, or the Spring 2013.

size of the reference group: put *White* together with other minority race/ethnicity groups as the reference group. This is also desirable because Hispanics is a group that is interested to the University. Since there were 54.41% Hispanic students and 27.36% whites (Table 5.10)

in our data, and we would like to know how Hispanic students do in Intermediate Algebra comparing to white Caucasian.

The p-value for the Chi-Square test is quite big (0.9771), and there is still large condition index after we put whites into the reference group, and this suggests that there is still some multicollinearity between the predictor variables. Remember there were only 4.43% of GED holders in the combined data (Table 5.9), and Table 5.3 on Spring data suggests that the pass rate is not significantly different between high school graduates and GED certificate holders. The backward elimination procedure also deleted *High school* from a logistic regression model at the 0.1 significance level. Thus, we decided to take away *High school* and *White* from the model. See Table 5.13, the variance inflation factors are all smaller than 1.2, and the biggest condition index is 20.857. The model should be free of multicollinearity problem with these selected variables.

Variable	Parameter Estimate	Standard Error	t Value	$Pr > \|t\|$	Tolerance	Variance Inflation
Semester GPA	0.12654	0.01314	9.63	$< .0001$	0.89919	1.06187
Course load	-0.00303	0.00405	-0.75	0.4549	0.85474	1.11211
SAT/ACT	0.16196	0.02538	6.38	$< .0001$	0.90015	1.16994
HS grad. years	-0.02537	0.00465	-5.46	$< .0001$	0.89352	1.11093
redesign	-0.13108	0.01152	-11.38	$< .0001$	0.86126	1.16108
Hispanic	0.0149	0.02404	-0.68	0.499	0.87106	1.14803
native American	-0.04278	0.04412	-0.97	0.3324	0.89939	1.11187

Table 5.13: Multicollinearity dianostics. "redesign" is a variable telling wether a student took Intermediate Algebra in the Redesign in Fall 2012, Traditional lecture in Fall 2012, or the Spring 2013.

We add three interaction terms: SAT/ACT * $HS\ grad.\ years$, $HS\ grad.\ years$ * $Hispanic$, and SAT/ACT * $Hispanic$, because the results from backward elimination suggest that these three interaction terms are significant.

5.2.4 Multilevel Logistic Model

We are curious about what happens if we fit the data with a standard multilevel logistic model. Note that, the response variable of this model has to be binary, so we create a binary response variable: pass or fail, with the letter grades "I", "AUD", and "W" as well as "C-", "D"(with +/-), "F" and "NC" being treated as fail. Table 4.13 and Table 5.15 are the fix and random effects for the multilevel logistic model:

$$p_i = logit^{-1}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau})$$

Where p_i is the probability of student i passing Intermediate Algebra, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design vectors for the fixed and random effects, respectively. In our case, $\boldsymbol{X}_i$ is the value for a list of variables about students' information, and $\boldsymbol{Z}_i$ is a 12×3 matrix consists of 0s and 1s telling if the student took Intermediate Algebra in Spring 2013, the pilot Redesign in Fall 2012 or a Traditional lecture section in Fall 2012. $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are fixed and random effects.

The test of covariance parameters based on the residual pseudo-likelihood gives a p-value $< .0001$, this suggests that the difference in the random intercepts is not 0, in other words, there is significant difference in students' probability to pass the course between students taking the course in Spring 2013, in the pilot Redesign of Fall 2012 and in the Traditional lecture sections of Fall 2012. Table 5.14 and 5.15 provide a description of the estimates for the fixed effects and random effects of the multilevel logistic model. We see that the random intercept for the lecture sections in Fall 2012 is the highest, then the pilot Redesign in Fall 2012, the intercept for the Spring Redesign is the lowest. This is likely caused by treating the over 50% of Incompletes (I) and Audits (AUD) in the Redesign of both Spring 2013 and Fall 2012 as a failing grade.

variable	Estimate	Standard Error	t	$Pr > \|t\|$
Intercept	-1.2869	0.5476	-2.35	0.0501
semester GPA	0.6897	0.07498	9.20	$< .0001$
Course load	-0.01063	0.02049	-0.52	0.6041
SAT/ACT	1.1707	0.2288	5.12	$< .0001$
HS. grad. years	-0.09345	0.03682	-2.54	0.0112
Gender	-0.1649	0.1055	-1.56	0.1183
Hispanic	-0.2192	0.1833	-1.20	0.2319
Native American	-0.1691	0.2271	-0.74	0.4566
SAT/ACT* HS grad. years	-0.03309	0.04592	-0.72	0.4713
HS grad. years*Hispanic	-0.1604	0.05769	-2.78	0.0055
SAT/ACT*Hispanic	-0.6739	0.2652	-2.54	0.0111

Table 5.14: Estimates of the fixed effects. The column "variable" lists the name of variables we used for the model, and the variable "Gender" takes female as reference group. The column "Estimate" displays the parameter estimate for variables listed in the first column with the multilevel logistic model; the column "Standard Error" gives the standard error of the parameter estimate in the multilevel logistic model, the column "t" provides the value of t-statistic and $Pr > |t|$ gives of the p-value for each parameter estimate.

Effect	Subject	Estimate	t	$Pr > \|t\|$
Intercept	Lecture	0.7742	1.93	0.1887
Intercept	Redesign	-0.3639	-0.89	0.4596
Intercept	Spring	-0.4103	-1.02	0.4108

Table 5.15: Estimates of the random effects. In the column "Subject" , "Lecture" means the Traditional lecture section in the Fall semester of 2012, "Redesign" represents the Redesign in Fall 2012, and "Spring" means the Redesign in the Spring semester of 2013. "Estimate" column gives the parameter estimates for its left column; the last two columns provides the t-statistic and p-value for each parameter estimate.

Prediction with the multilevel logistic model

Using the parameter estimates of the multilevel logistic model in Table 5.14 and Table 5.15, for any student from Traditional lecture sections in the Fall semester of 2012, the estimated

probability of passing the course is estimated as:

$$\begin{aligned} logit^{(-1)}(\hat{p_i}) &= (-1.2869 + 0.7742 + 0.6897 \times semesterGPA - 0.0106 \times Courseload + 1.1707 \\ &\quad \times SAT/ACT - 0.0935 \times HS\ grad.\ years - 0.165 \times male - 0.2192 \times Hispanic \\ &\quad -0.169 \times Native\ American - 0.0331 \times SAT/ACT \times HS\ grad.\ years - 0.1604 \\ &\quad \times HS\ grad.\ years \times Hispanic - 0.6739 \times SAT/ACT \times Hispanic) \end{aligned}$$

To estimate the probability of passing the course for a student from the pilot Redesign in Fall 2012, or in Spring 2013, simply change the random effect coefficient 0.7742 to -0.3639 or -0.4103. We left the non-significant terms in the model for comparison's sake.

Suppose there is a student with average *semester GPA* (2.863), *Course load* (13.862), SAT/ACT (-0.561) and $HS\ grad.\ years$ (2.815), who is a Hispanic male, and took Intermediate Algebra in Fall 2012 in one of the Traditional lecture sections, his probability of passing Intermediate Algebra would be 53.4%. If he was in the pilot Redesign last Fall, then his probability of passing the course is 26.8%; if he took the course in Spring 2013, his probability of passing the course would drop to 25.9%. If the student is a Hispanic female, her probability of passing Intermediate Algebra in the lecture section in Fall 2012, the pilot Redesign in Fall 2012 and the Redesign in Spring 2013 would be 53.4%, 27.2% and 26.3%, respectively.

This predicted probabilities are very low because all other grades except A, B (with +/-) and $C+$, C, CR are treated as a fail. Since there were over 50% Incompletes in the Redesign, if treating the Incompletes as a a pass, the pass rate will be extremely high. Either way, the prediction from the multilevel logistic model will be far from the truth and is not reasonable. Moreover, it gives only a probability of passing and ignore the difference among the passing letter grades, and the difference among the failing letter grades. This brings the necessity to apply our new models to analyze the data to make a better use of the grading information, and these models consequently provide a more precise prediction of the student success.

5.2.5 Analysis with the Bayesian Partially Ordered Multinomial Probit and Logit Models

As we have seen in the last section, the multilevel logistic model in Section 5.2.4 distorts the more than 50% of Incomplete and Audit grades in the Redesign, and utilizes only part of the information conveyed by the letter grades. It is not an accurate measurement to treat the Incomplete grades as either fail or pass, since these students still have a chance to complete the course. Forcing the response variable to be binary throws away the original grading information we have. To overcome the shortcomings of traditional multilevel modeling and to use all the information in the data, we apply our Bayesian partially ordered probit and logit models described in Chapter 4 to compare student learning performance between the Redesign in Spring 2013 and the pilot Redesign in Fall 2012, as well as the Redesign in Spring 2013 versus the Traditional lecture sections in Fall 2012.

For the sake of comparison, we include all variables and interaction terms from the multilevel logistic model in Section 5.2.4. In Chapter 4, we compared the student performance between Traditional lecture sections and the pilot Redesign by comparing the students performance taught by all the instructors, because there was one instructor who taught only the Redesign, and other instructors only taught the Traditional lecture sections. Now we lump all instructors who taught Traditional lecture sections in Fall 2012 together and all instructors teaching the Redesign in Spring 2013 together. This is because we mainly care about the comparison between Traditional lecture sections and the Redesign.

We ran the algorithm for 100,000 iterations, kept every 10th observation (thinning) to save space and burned out the first 1/4 observations to reduce the dependency of the sample on the initial values. Figure 5.3 displays the trace plots for several parameters in the model on the combined Fall and Spring data. The plots suggest that there is no evidence of serious departure from the stationary. We omitted the remaining trace plots for brevity.

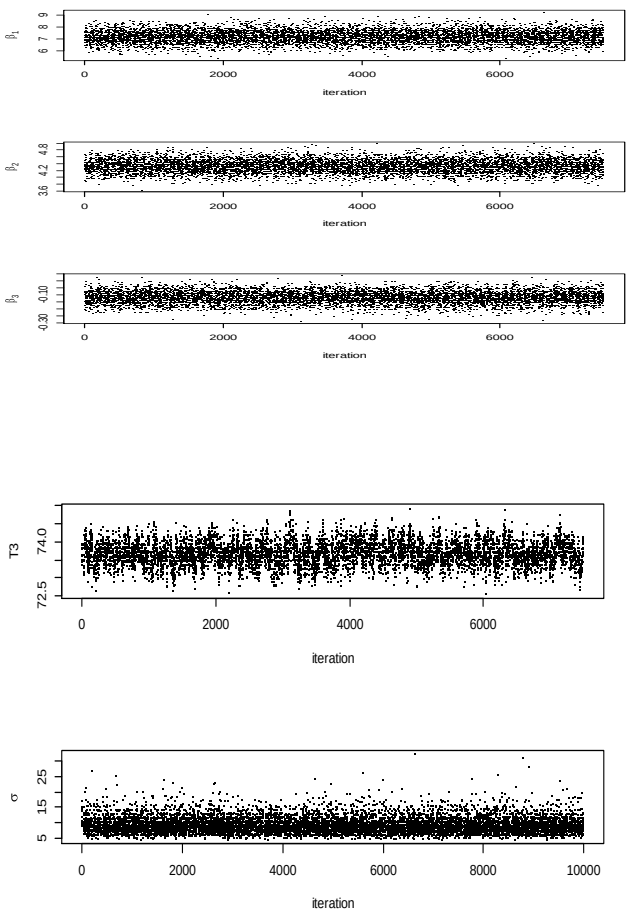

Figure 5.3: Trace plots for convergence diagnostics. The first three plots are the posterior draws for β_1, β_2 and β_3, respectively. The last two plots are the posterior draws for τ_3 and σ_ϵ.

Estimated posterior means for the coefficients of Bayesian partially ordered probit and logit models

Table 5.16 lists the parameter estimates for each element of $\boldsymbol{\beta}$. The estimated coefficients for the probit and the logit model have the same sign, and are close in magnitude, so the interpretation of them are similar. We focus on the probit model for the rest of this section.

The estimated coefficients for SAT/ACT and *Semester GPA* are positive and big, which mean that students who did well in SAT and ACT math, or have a higher GPA in the semester, tend to do better in Intermediate Algebra. Specifically, the estimated coefficient of SAT/ACT means that if the standardized SAT/ACT (ranges from -4 to 4) increases one point, the student's course score (in a range of 0 to 100) will increase on average 7.197 points while controlling other variables. If *Semester GPA* (in a range of 0 to 4) increases one point, the course score will tend to increase 4.327 points, while holding other variables constant. *Course load* has a negative coefficient, indicating that students who took more courses tended to do worse in Intermediate Algebra. Fixing the values of other variables, if a student takes one more credit hour, his or her average course score in Intermediate Algebra will likely decrease by 0.118 points, so if a student takes one more course of 3 credit hours, his or her score in Intermediate Algebra will decrease for 0.354 points.

The coefficient for *HS grad. years* (high school graduation years) is negative, meaning that students who wait a longer time to take Intermediate Algebra after they graduated from high school tend to do worse. Marginally, if a student's time of high school graduation by the time of taking Intermediate Algebra is one year longer, his or her average Intermediate Algebra course score will go down for 0.326 points.

Gender is set to 0 for male and 1 for female, so the positive coefficient for gender means that female students did better than male students, on average, the score of female students is 1.53 points higher than male students. *Hispanic* has a small positive coefficient, so it has a positive effect on Intermediate Algebra course score, suggests that Hispanic students might

do a little better than students in the reference group. However, $Hispanic$ has an interaction with $HS\ grad.\ years$ and SAT/ACT with an estimate -0.106, which increases the negative effect of $HS\ grad.\ years$ and decrease the positive effect of SAT/ACT score on Intermediate Algebra success for Hispanic students. Thus, we cannot say if Hispanic students did better than other race/ethnicity groups overall. The negative coefficient of native American tells that that native American students did a little bit worse than other race/ethnicity groups, and they had on average 0.532 points lower than other race/ethnicity groups on average with the same values of other variables.

The negative coefficients (-0.106 for the probit model, -0.144 for the logit model) of the interaction term SAT/ACT * $HS\ grad.\ years$ means that the longer a student graduated from high school (or earned their GED certificate), the less the effect of the SAT/ACT score is on students' Intermediate Algebra grade. This is reasonable, since most students take their SAT/ACT in the junior or senior year of high school, the longer the $HS\ grad.\ years$ is, students tend to forget more of the math knowledge they learned in high school, which also was measured by the SAT/ACT test, so the effect of SAT/ACT on students' Intermediate Algebra success is decreased. Similarly, the higher the SAT/ACT score is, the less the effect of $HS\ grad.\ years$ on Intermediate Algebra score is. This is because the better students did in the SAT/ACT tests, the less they tend to forget after the same amount of $HS\ grad.\ years$, the effect of $HS\ grad.\ years$ on students' Intermediate Algebra score tend to decrease with SAT/ACT getting higher.

$HS\ grad.\ years$ * $Hispanic$ have a coefficient -1.025 (or -0.969 for the logit model), which means that the coefficient for $HS\ grad.\ years$ we just saw (-0.326) is for the race/ethnicity groups other than Hispanics, and for Hispanic students, it should be -0.326 -1.025 = -1.351 for the probit model, and -0.319-0.969= -1.288 for the logit model. In other words, for Hispanic students, the effect of $HS\ grad.\ years$ on Intermediate Algebra grade is worse (negative and at a bigger magnitude). For the same value of $HS\ grad.\ years$, Hispanic students' grades tend to be lower. The coefficients of SAT/ACT * $Hispanic$ (-1.868 or

-2.075) mean that the effect of SAT/ACT for Hispanic students on Intermediate Algebra course score should be lower, and the coefficients should be: 7.197-1.868 = 5.329 for the probit model, this indicates that for the students with the same SAT/ACT score, Hispanic students tend to do a little worse than students of other race/ethnicity groups on average. Again, these results are consistent with education literature.

	Probit	logit
SAT/ACT	7.197	7.654
Semester GPA	4.327	4.319
Course load	−0.118	−0.115
HS grad. years	−0.326	−0.319
Gender	1.530	1.648
Hispanic	0.226	0.205
Native American	-0.532	−0.470
SAT/ACT * HS grad. years	-0.106	−0.144
HS grad. years × Hispanic	-1.025	−0.969
SAT/ACT × Hispanic	-1.868	−2.075

Table 5.16: the column under "Probit" gives the posterior mean of the elements of $\boldsymbol{\beta}$ (the estimated coefficients for each predictor variable and the interaction terms) obtained with the Bayesian partially ordered probit model, and the column under "Logit" lists those obtained by the Bayesian partially ordered logit model.

Figure 5.4 displays the boxplots of all the parameters estimated from the sample for each element of $\boldsymbol{\beta}$, the rightmost three boxplots are for the interaction terms: SAT/ACT * $HS\ grad.\ years$, $HS\ grad.\ years$ * $Hispanic$ and SAT/ACT * $Hispanic$.

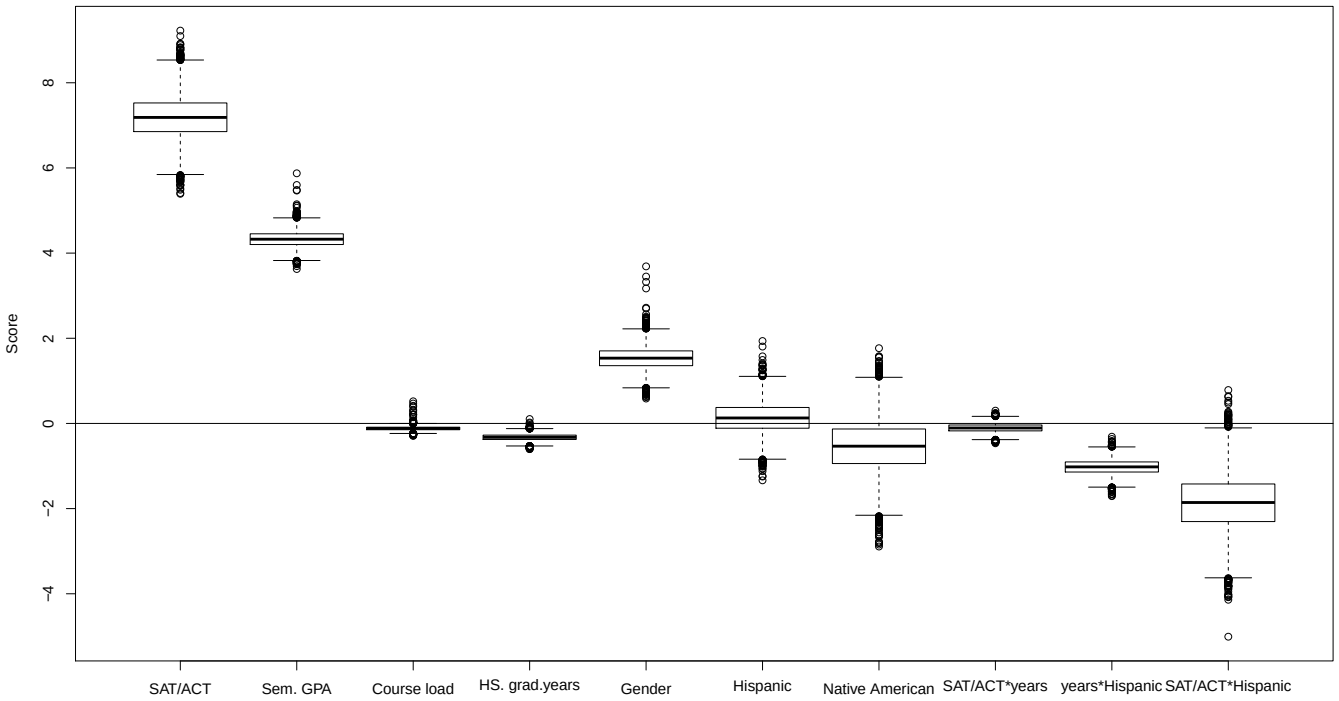

Figure 5.4: Marginal posterior distributions of the estimated elements of $\boldsymbol{\beta}$. The solid horizontal line represents the overall mean. $Sem.GPA$ represents "Semester GPA": a weighted average of students' GPA in Spring 2013 excluding the course of Intermediate Algebra, *years* is brief representation of $HS\ grad.\ years$ for saving space on the graph.

Table 5.17 gives the estimated posterior means for the random effects $\boldsymbol{\tau}$ for students taking the course in Traditional lecture sections or the pilot Redesign in Fall 2012, as well as in the Redesign Spring 2013. We see from Table 5.17 that the estimated intercept for the pilot Redesign in Fall 2012 was the highest, the 2nd was the Redesign in Spring 2013, and it was the lowest for the lecture sections in Fall 2012. The random effect serves as an intercept for the probit and logit models, that is, the models to predict student course score from the Fall 2012 Traditional lecture section, the Fall 2012 pilot Redesign and the Spring 2013 Redesign have different intercepts. The difference of the posterior mean of the random intercepts for the Redesign in Fall 2012 and Spring 2013 is only about 2 points (73.67 and 75.54), but they are all much higher than that of Traditional lecture sections in Fall 2012 (67.06). This difference indicates that the students' course score in the Fall and Spring Redesign, as well as Traditional lecture sections are different. The average quantity of difference is: about 2-point difference between the two Redesigns, and the Spring Redesign has 6.5 points higher than the Traditional lecture in Fall 2012 on average.

Figure 5.5 gives a boxplot of all random effects for the course section categories estimated from the posterior draws. The estimated random intercepts for the Redesign in Fall 2012 and Spring 2013 are significantly higher than those in the Traditional lecture sections in Fall 2012, while there is an overlapping between the Redesigns in Fall 2012 and Spring 2013. Note that there is a large variance for the Redesign in Fall 2012, which is caused by the small sample size: we had only 216 students, while there were four times more students in the Redesign in Spring 2013 (950 students). Thus, we conclude that the difference between the Redesign in Fall 2012 and Spring 2013 is not significant.

Section	Probit	Logit
Traditional Lecture in Fall 2012	67.057	66.805
The Pilot Redesign in Fall 2012	75.544	75.989
The Redesign in Spring 2013	73.671	73.303

Table 5.17: The estimated posterior means of the random effects under the probit model and logit model. The column under "Section" provides the information of students taking Intermediate Algebra in the Traditional lecture sections in Fall 2012, the pilot Redesign in Fall 2012, or Redesign in Spring 2013. The column under "Probit" are random effects estimated from the partially ordered probit model, and the column under "Logit" means the random effects estimated from the partially ordered logit model.

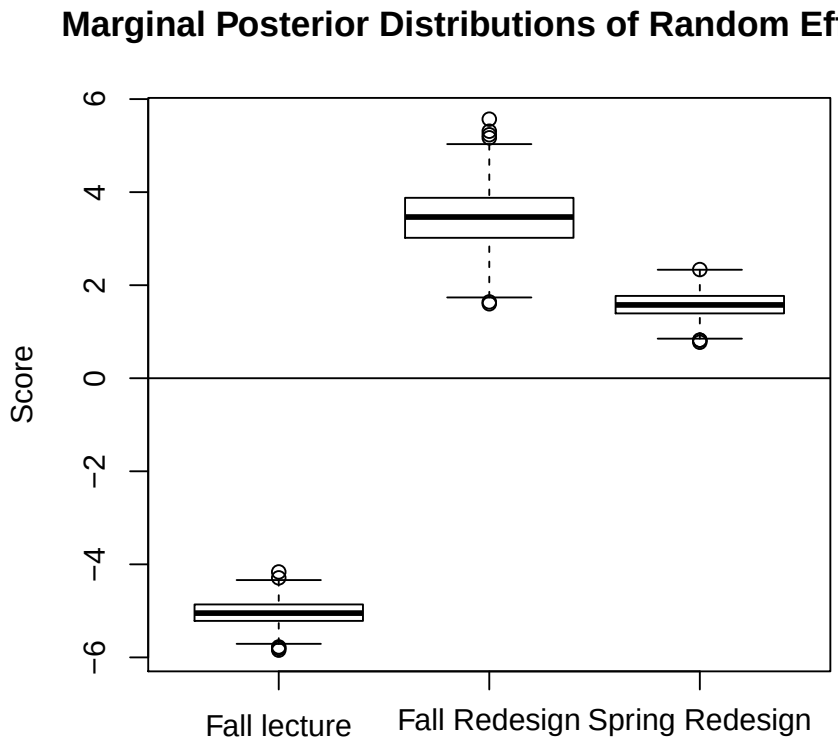

Figure 5.5: A boxplot of random effects ($\boldsymbol{\tau}$s). "Fall lecture" describes the the distribution of posterior draws for the Traditional lecture sections in Fall 2012; "Fall Redesign" gives the distribution of posterior draws for the pilot Redesign in Fall 2012, and "Spring Redesign" displays the distribution of posterior draws for the Redesign in Spring 2013.

Prediction with Bayesian partially ordered probit and logit models

Once the posterior parameter draws are obtained, we can use them to predict the course score of students. The University is very diverse, but one race/ethnicity of interest is Hispanic. Hispanic students comprise 38.22% of the student body (UNM, 2013), about the same percentage as whites in the University. About 50-60% of Intermediate Algebra students are Hispanic. Thus, we care about how Hispanic students do in this course. For a Hispanic male with an average SAT/ACT score (SAT Mathematics 451 or ACT Math 18), average *Semester GPA* (2.863), *Course load* (14 credit hours excluding Intermediate Algebra) and average HS *grad. years* (2.815 years), Figure 5.6 displays the posterior distribution of the Hispanic student's predicted scores estimated from the Bayesian partially ordered probit model if he was taught by the Traditional lecture method in Fall 2012, the Redesign in Fall 2012, or by the Redesign in Spring 2013, respectively. Figure 5.7 describes the posterior distribution of the predicted scores by the Bayesian partially ordered logit model if the student was taught by the lecture in the Fall 2012, the Redesign in Fall 2012, and the Redesign in Spring 2013.

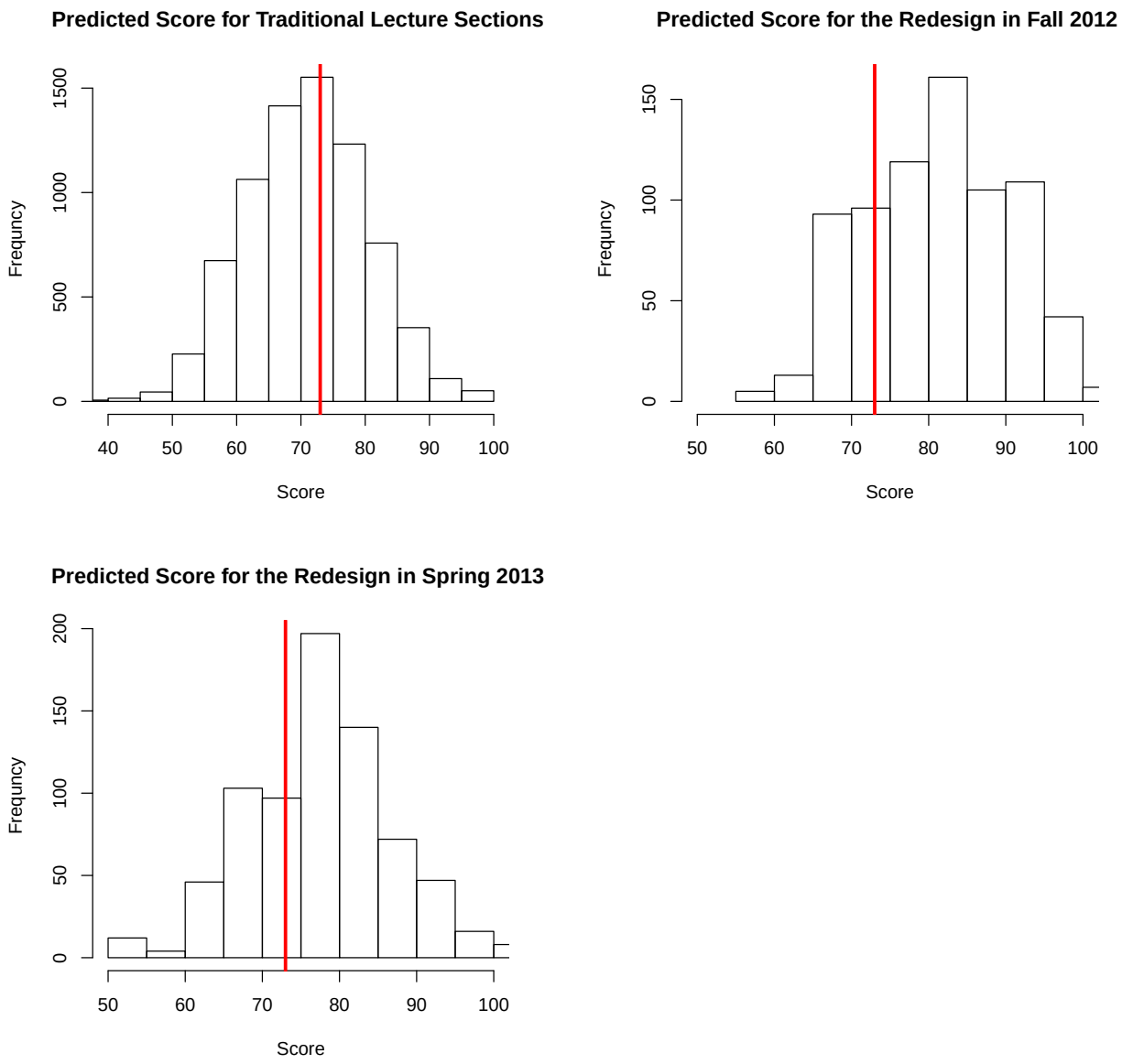

Figure 5.6: Distribution of predicted scores by the partially ordered probit model for a Hispanic male with average information. Left upper panel: the student took the course in the Fall semester of 2012 by the Traditional lecture method. Right upper panel: the student took the course in the Redesign in Fall 2012. Lower panel: the student took Intermediate Algebra in the Redesign in Spring 2013. The red line is the passing score (73).

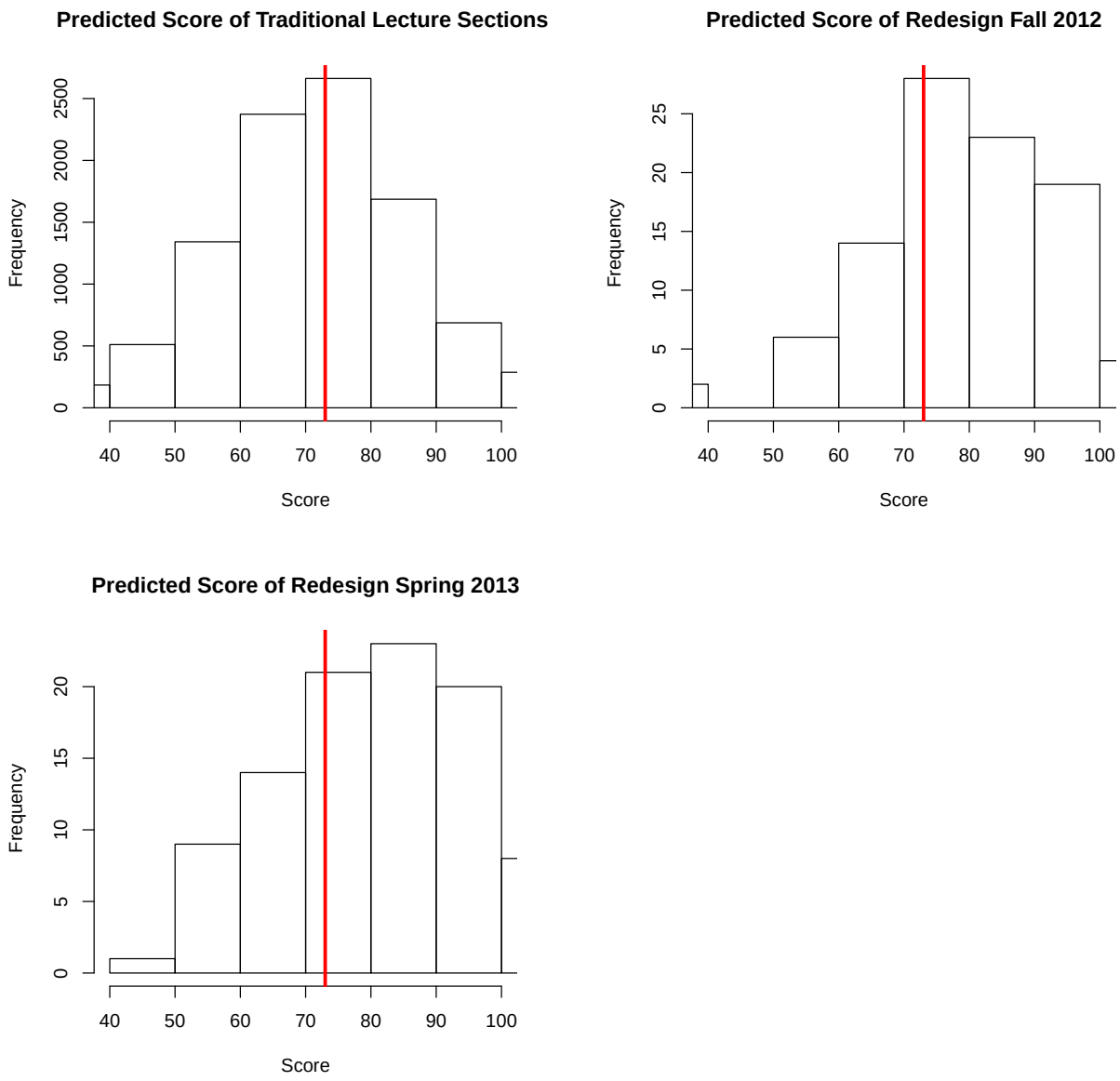

Figure 5.7: Distribution of predicted scores by the partially ordered logit model for a Hispanic male with average information. Left upper panel: the student took the course in the Fall semester of 2012 by the Traditional lecture method. Right upper panel: the student took the course in the Redesign in the Fall, 2012. Lower panel: the student took Intermediate Algebra in Spring 2013. The red line is the passing score (73).

Table 5.18 lists the mean of the predicted score and predicted probability of passing the course for the example Hispanic student estimated from both our Bayesian partially ordered probit and logit model. If he took Intermediate Algebra with the Traditional lecture method in the Fall semester of 2012, his average score is 71, and his probability of passing the course is 0.44; if he was in the Redesign in the Fall semester of 2012, his score would be around 80, his probability of passing the course is about 0.75. His score and passing probability would be a little lower if he took it in the Redesign in Spring 2013, about 78 by the probit model, 77 by the logit model, and the passing probability is around 0.71. On average, students' score is about 7 points (2/3 of a letter grade) higher if they switch from Traditional lecture sections to the Redesign. There is not a significant difference in students' score between the Redesign in Fall 2012 and Spring 2013 (only about 1.5 points on average).

Section	Score		Probability of Pass	
	Probit	Logit	Probit	Logit
Traditional lecture in Fall 2012	71.43	71.4	0.44	0.44
The pilot Redesign in Fall 2012	79.28	80.76	0.75	0.75
The Redesign in Spring 2013	77.99	77.98	0.71	0.71

Table 5.18: The predicted score and passing probability for the Hispanic male with both the Bayesian partially ordered probit and logit models. The two columns under "Score": Probit and Logit provide the predicted score by the partially ordered probit model and logit model, respectively. The two columns under "Probability of Pass" gives the predicted probabilities by the partially ordered probit or logit model, respectively.

Overall, according to the results from the Bayesian partially ordered probit and logit model, we conclude that students' performance measured by the course score and the pass rate in the lecture sections in Fall 2012 is the lowest, students' performance in the Redesign in Fall 2012 and the Spring 2013 is significantly higher than that of the Traditional lecture sections. Students' course score and pass rate in the full Redesign in Spring 2013 are a little bit lower than that of the pilot Redesign in Fall 2012, and this difference is not significant.

Hence, we have evidence that the Redesign is better than the Traditional lecture method in terms of students' learning performance.

These results obtained with the partially ordered probit/logit models are based on the assumption that the Incomplete students in the Redesign have the same situation as those in the Traditional lecture sections. We know that instructors rarely give an "I" in a Traditional lecture class in the University, However, in the Redesign, the grade "I" may related to students' math ability, and these students may be less motivated in a self-paced environment. Of course there might be also some students in the Redesign who did not finish the course within the semester because some trauma happened to them in the middle of the semester.

5.3 Analysis on Students' Continuing Success in Higher Level Math Courses Following the Fall Semester of 2012

Recall that our primary research questions for the evaluation component were:
"What is the impact of the Intermediate Algebra course redesign on:

1. Student success in Intermediate Algebra in the semester taken, and

2. Student success in more advanced mathematics courses (such as College Algebra, A Survey of Mathematics, and Introduction to Statistics) taken in the subsequent semesters?"

To answer the second question, we acquired course and grade data in the Spring semester of 2013 for all students who registered for Intermediate Algebra in Fall 2012 after the third week of the semester.

There were 1115 students who enrolled in Intermediate Algebra in the Fall 2012 and continued to take courses in the Spring at the University. But only 586 students who took Intermediate Algebra in Fall 2012 continued to take College Algebra, A Survey of Mathematics, or Introduction to Statistics in Spring 2013. Recall that we had grades for 1308 students from the Fall 2012, thus, 193 of them did not take any course in the Spring 2013 at the University after they took Intermediate Algebra in the Fall 2012. Further information about these 193 students can be collected, for example, will some of them come back to the University some time later? Have some of them transferred to other universities? What are these students doing right now? We suspect part of these students might drop from the college entirely after failing Intermediate Algebra several times. Unfortunately, this frequently happened to students who attempted Intermediate Algebra several times and failed, which means that they cannot earn their degree at the University. They became frustrated and dropped out of college. The negative impact of this phenomenon to the students and to the country was discussed in Section 2.1. We merged the Spring higher level math course grade data with the cleaned Intermediate Algebra data in the Fall 2012. This gives us the data for the 586 students who took Intermediate Algebra in the Fall semester of 2012 and continued to take College Algebra, A survey to Mathematics, or Introduction to Statistics, after they finished Intermediate Algebra in Fall 2012.

According to our analysis on the Fall 2012 Intermediate Algebra grade data, 744 students passed Intermediate Algebra in the semester. Now we know that 586 students among them continued to take upper level math courses College Algebra, A Survey of Mathematics, and Introduction to Statistics. The remaining 158 students will take one of these three courses later to fulfill the requirement of their degree program.

Table 5.19 gives the distribution of students in the upper level mathematics courses in the semester of Spring 2013. There were 370 (63.14%) students taking College Algebra, 133 (22.70%) students took Introduction to Statistics, and 83 (14.16%) students took A Survey of Mathematics in the Spring 2013.

Courses	Frequency	Percent	Cumulative Frequency	Cumulative Percent
College Algebra	370	63.14	370	63.14
Introduction to Statistics	133	22.70	503	85.84
Survey of Mathematics	83	14.16	586	100.00

Table 5.19: Distribution of students who took Intermediate Algebra in the Fall 2012 and upper level math courses during the Spring semester of 2013.

Comparisons of the pass rate of students from Traditional lecture sections and Redesign in higher level math courses

A comparison of overall pass rate For all the 586 students who continued to take higher level math courses in the Spring semester of 2013, Table 5.20 gives the pass rate of students in the three higher level courses in the Spring who took Intermediate Algebra in the Traditional lecture sections is 68.48%, and the pass rate in the three higher level math courses in the Spring for Intermediate Algebra Redesign students is 69.83%, about the same. The p-value 0.988 from the Chi-square test suggests that there is no difference in the pass rate in the upper level math courses between students who took Intermediate Algebra from the Traditional lecture sections and those who were from the Redesign in Fall 2012.

Comparisons of the pass rate in each class We divide the data into three subsets: students in College Algebra, A Survey of Mathematics and Introduction to Statistics to investigate the pass rate in each course between students from the Intermediate Algebra Traditional lecture sections and those from the Redesign.

A comparison of the pass rate in College Algebra Table 5.21 describes that the pass rate in College Algebra between students coming from the Intermediate Algebra Traditional lecture sections and the Redesign in Fall 2012 are 64.18% and 65.71%, respectively. The p-value from the Chi-square test is 0.8568, which suggests that the pass rate in College

	Traditional lecture	Redesign	Total
Passed students	365	37	402
Passed percentage among all students	62.29	6.31	68.60%
Pass rate among each group	**68.48**	**69.81**	
Failed students	168	16	184
Failed percentage among all students	28.67	2.73	31.40%
Failing rate in each group	31.52	30.19	
Total number of students in each group	533	53	586
Percentage of students in each group over all	90.96	9.04	100.00

Table 5.20: Overall pass rate of the 586 students in continuing upper level math courses between those from Intermediate Algebra Traditional lecture sections and Redesign. The "Passed percentage among all students" means the percentage of passed students from each Intermediate Algebra teaching method over all 586 students enrolled continued to take higher level math courses in Spring 2013. Similarly, the "Failed percentage among all students" means the percentage of failed students from each Intermediate Algebra teaching method over all 586 students enrolled continued to take upper level math courses in Spring 2013. The "Pass (Failing) rate" represents the percentage of passed (failed) students in the higher level courses from each Intermediate Algebra teaching method over all student in each group (taught by the Traditional lecture or the Redesign).

Algebra between students taking Intermediate Algebra from Traditional lecture sections and the Redesign is not significantly different.

A comparison of the pass rate in Introduction to Statistics Table 5.22 displays students passing information in Introduction to Statistics between those who took Intermediate Algebra in the Traditional lecture sections and Redesign. Among the 133 students who took Intermediate Algebra in Fall 2012, then Introduction to Statistics in Spring 2013, 97 of them passed the course. 72.80% of the students were from the Intermediate Algebra lecture

College Algebra	Traditional lecture	Redesign	Total
Passed students	215	23	238
Passed percentage among all students	58.11	6.22	64.32%
Pass rate among each group	**64.18**	**65.71**	
Failed students	120	12	132
Failed percentage among all students	32.43	3.24	35.68%
Failing rate in each group	35.82	34.29	
Total number of students in each group	335	35	370
Percentage of students in each group over all	90.54	9.46	100%

Table 5.21: Pass rates in College Algebra between students taking Intermediate Algebra from Traditional lecture sections and the Redesign.

sections, and 75% of them from the Redesign. The Chi-Square test gives the p-value 0.95, suggesting that there is no difference in Introduction to Statistics pass rate between students who took Intermediate Algebra by the Traditional lecture method and the Redesign in Fall 2012. There were only 8 students from the Redesign in Fall 2012 continued take Introduction to Statistics in Spring 2013, and 2 of them failed the course. One thing of concern is that the sample size of students in Introduction to Statistics from the Redesign is too small.

A comparison of the pass rate in A Survey of Mathematics Table 5.23 displays students' passing information in A Survey of Mathematics between those who took Intermediate Algebra in the Traditional lecture sections and those in the Redesign. Among the 83 students who took Intermediate Algebra in Fall 2012 and continued to take A Survey of Mathematics in the following Spring, 73 of them took Intermediate Algebra from the Traditional lecture sections and there were only 10 students taking Intermediate Algebra in the Redesign. 80.82% of the students from the Intermediate Algebra Traditional lecture

Introduction to Statistics	Traditional lecture	Redesign	Total
Passed students	91	6	97
Passed percentage among all students	68.42	4.51	72.93%
Pass rate among each group	**72.80**	**75.00**	
Failed students	34	2	36
Failed percentage among all students	25.56	1.50	27.07%
Failing rate among each group	27.20	25.00	
Total number of students in each group	125	8	133
Percentage of students in each group over all	93.98	6.02	100%

Table 5.22: Pass rates in Introduction to Statistics between students coming from Intermediate Algebra Traditional lecture sections and the Redesign.

sections passed A Survey of Mathematics and 80% of them were from the Redesign in Fall 2012 passed. The p-value of the Chi-Squire test is 0.95, suggesting that there is no difference in Survey of Mathematics pass rate between students who took Intermediate Algebra from Traditional lecture sections and those from the Redesign in the previous Fall semester. Again, we are concerned that the students in Survey of Mathematics and from the Redesign is too small.

A Survey of Mathematics	Traditional lecture	Redesign	Total
Passed students	59	8	67
Passed percentage among all students	71.08	9.64	80.72%
Pass rate among each group	**80.82**	**80.00**	
Failed students	14	2	16
Failed percentage among all students	16.87	2.41	19.28%
Failing rate in each group	19.18	20.00	
Total number of students in each group	73	10	83
Percentage of students in each group over all	87.95	12.05	100%

Table 5.23: Pass rates in A Survey of Mathematics between students who took Intermediate Algebra in Traditional lecture sections and Redesign.

Chapter 6

Discussion and Future Work

6.1 Discussion

Student success is often measured using letter grades, which are sometimes totally ordered but often have only a partial ordering. The Bayesian partially ordered probit and logit models presented in this dissertation are a way of retaining all of this ordering information. These models are generalizations of the standard models in the sense that binary or ordered response variables, which are required for the standard models, are special cases of partially ordered response variables. We used a Bayesian approach because prior knowledge about the parameters can be incorporated into the model and the Bayesian computation tools can be used to fit the models. Additionally, the model we present is understandable to non-statisticians. The motivation for these models was evaluating the impact of a course redesign. In this example, we observe students' letter grades and many students received Incomplete, the analyses from the standard models are strongly biased because those models can not use well the unordered information. Applying our partially ordered probit and logit models to the data, we could use all the grading information to predict students' course score, and compare students' success in the Traditional lecture sections and the Redesign.

Students' background and instructor information were also utilized through covariates.

With the Fall 2012 data, the results from our models show that the students from the Redesign scored higher and consequently have a better probability to pass the course than those from the Traditional lecture sections on average, and the application on the combined data of Fall 2012 and Spring 2013 also suggests that the Redesign in the two semesters both the Fall and the Spring was superior to the Traditional lecture teaching method, basing on the predicted students' performance with our models. The different random effects, the predicted course score and passing probability of students by the models quantify part of the benefits students gain by switching from the Traditional lecture sections to the Redesign. Thus, we have evidence that the Redesign is effective in helping students learn Intermediate Algebra.

The partially ordered probit model outperforms totally ordered and binary probit models in terms of cross-validated mean squared error and area under the ROC curve. This is not surprising since it includes more information (the partially ordered subjects) in the same general model structure. In a sense, the model automatically performs an imputation on the partially ordered subjects. The gain we see is thus related to the gain in modeling the "missing" data.

There are, however, limitations to the study and clear areas of needed future work. In our example, the model has the potential to help fill in the missing data represented by the large number of Incompletes; however, the method we used assumes that Incompletes were randomly assigned to students, when this may not be the case. While it is true that if a student gets an Incomplete in a Traditional lecture section where such grades are reserved for extreme situations like an illness or family emergency, it may not be the case in a self-paced course that builds such grades into its structure. In other words, it is reasonable to assume that the Incomplete grades are randomly assigned in a Traditional lecture section, but they may reflect something about a student's study skills, for example, in the Redesigned course. Settling this issue will require examining the actual score that students eventually receive in

the course. This data was not available at the time of the study.

We developed this model specifically for the Redesigned course evaluation discussed in Section 3.1.1, but it is easily extended to other categorical data problems with a partially ordered response variable. For example, criminal sentences are partially ordered: one year in prison is clearly less than two years. But one year of prison with 3 years probation is not necessarily less than 16 months in jail. A data set containing sentences, crimes, and demographic characteristics of the criminals could be analyzed with this model. Another example is compensation and benefits in job offers: a job with a yearly payment \$100,000 is clearly better than a job paid with \$80,000 per year, but if the $\$80,000$ provides two months paid vacation and good benefit while the \$100,000 one does not, we can not say which one is better; what if another job paying $\$70,000$ with three months paid vacation, flexible schedule and even better benefits? There is no clear order that one job is better than another. If there is a dataset containing partially ordered job categories according to salary, benefits, paid vacation, stability of the job, etc, with people's demographic information, we could apply the partially ordered probit and logit models to predict people's job preference. This could help the employers set offer strategies to attract the best employees they want. Other partially ordered examples such as different situations of alcoholism: nondrinker, former drinker, current social drinker, and current heavy drinker; different psychiatric conditions: no anxiety, mild anxiety, anxiety with depression, severe depression and severe anxiety. The models also work for censored data in a public health, medical or biology field, where the event of interest of some individuals occur before they enter the study, or is only known to happen in an interval, and sometimes some individuals drop out of the study because of unexpected reasons. Our models can also be applied to this censored data and the missing response data are imputed automatically by the model.

In education field, the boundaries for the partially ordered response variable: students' letter grades are usually known, so the variance of the error terms σ_ϵ^2 does not have to be fixed to make parameters identifiable. When the model is applied to data in other fields,

the limits of the latent variable for each observed response categories (we called it γ) may be viewed as unknown. The work of McCulloch et al. (2000) on defining a prior for σ_ϵ^2 (or covariance matrix) to get fully identified parameters can be used for the partially ordered probit model.

6.2 Future Work

After this dissertation, there are several interesting problems we would like to tackle.

Modifications to the model in Equation 4.1 can easily be made. Say we observe student performance in the higher level math courses for which Intermediate Algebra is a prerequisite, autoregressive error terms could be utilized in this case to model students' performance in all future courses.

In Section 4.9, we fixed the value of parameters $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, σ_ϵ^2, and simulated $\boldsymbol{X}$ and $\boldsymbol{Z}$, then used the model expressed in Equation 4.1 to simulate the latent variable S_i, and map the the latent variable S_i with a partially ordered response variable Y_i. We repeated this process for the sample size $n = 500, 1000, 3000, 5000$ and $10,000$. Figure 4.4 shows that with the sample size increasing, the posterior estimates concentrate more on the true value, and vary less. We used 12 instructors, and each instructor teaches about 40 students on average, hence the estimates of $\boldsymbol{\tau}$ are based on these about 40 students taught by each instructor when the sample size n=500. When the sample size increases to 1000, 3000, 50000 and 10,000, while the number of instructors stays the same, there would be about 80, 240, 400, 800 students taught by each instructor. The estimates of $\boldsymbol{\tau}$ will based on many more observations, of course, are more accurate and closer to the true value. Considering the situation that when the sample size increases, the length of $\boldsymbol{\tau}$s also increases while fixing the class size, we suspect that the estimate of $\boldsymbol{\tau}$s will not get closer to the true value. This is worth further investigation.

The Gibbs sampler, in this case, requires us to sample from the truncated normal distribution. This sampling can be slow if many dependent observations are needed. We will produce faster algorithms for this sampling. We would also like to adopt an Exact Hamiltonian Monte Carlo algorithm for Truncated Multivariate Gaussians (Pakman and Paninski, 2013) to make the fitting faster.

We analyzed students' continuing success in higher level mathematics course in the Spring semester of 2013 after they passed the Intermediate Algebra course in the previous semester, Fall 2012, in Section 5.3, and this analysis is only based on the 586 students who continued to take higher level math courses in Spring 2013. There were only 216 students in the Redesign in Fall 2012, of which 66 students passed at the end of the semester, and out of those 53 passed students enrolled in the higher level math courses in Spring 2013. Most of students enrolled in Intermediate Algebra in Fall 2012 did not passed the course at the end of the semester and would probably still continue to work on it. There were also some students who passed Intermediate Algebra but did not take any math courses in the following semester. Thus, the current analysis is limited because of the small number of students from the Redesign in Fall 2012, which was the available data at the time of this analysis. For a complete evaluation, we will require data about students enrolled in the Intermediate Algebra course for a longer period of time (say, 5 years), and all the courses they take in each semester and the letter grades for the course. Then we can fit the data to a dynamic linear model to do a time series analysis.

54.63% of students received Incompletes in the Redesign in Fall 2012, and this percentage is 56.21% in Spring 2013. How long did it take for students to pass the course? Upon following up and collecting data about the Incomplete students, we could build a survival function and a hazard function to predict the average time students take to finish Intermediate Algebra, and investigate how the predictor variables such as students' SAT/ACT, semester GPA, course load, high school graduation years, gender, race/ethnicity influence the time that students needed to pass the course.

Since there are so many Incompletes, students use the computers and facilities in the Redesigned course to keep working on the course for over one semester, these students take spots in the course, and become a burden to the department. We need to try our best to help and stimulate these students to finish sooner. The department chair suggested a monetary stimulus: students have to deposit a certain amount of money when they register for this course, and they can get the money back if they pass the course at the end of the semester, if not, the deposit money will be forfeited. This might be a good idea to push students to work hard on this course. How much money should we ask students to deposit? If the amount is too small, say \$5, then students will not care about it. If we set the amount of money too big, suppose \$2000, then many students can not afford it, they would turn to other universities like the local community college to take Intermediate Algebra. We could design an experiment to find the optimum amount of money that is most effective in motivating students pass the course soon.

The context of this problem is to make a decision: Traditional or Redesign? Given the prior research about math course redesign from other institutions, the benefit to student achievement appears to be worth the cost of changing instructional methods from the Traditional lecture to the Redesign. Conditional on the data from the University and the model fit, do we come to the same decision? Simply said, what minimum improvement in student achievement is worth switching from Traditional to Redesign? We can develop a loss function to encapsulate the costs of Traditional and the cost of Redesign to help decide which is preferred given our evidence.

We wish to examine the impact of the prior distribution. Specifically, we wish to define "good" priors for repeated use over different data. We plan to perform a sensitivity analysis to understand the impact of the priors on the posterior distributions, including, in a simulation study, whether "good" priors are substantially better than "non-informative" priors and whether "unreasonable" priors greatly distort inference, and under which situations (such as for various sample sizes). The non-informative priors such as Jefferey or reference priors

have not been well-defined for multilevel (random effect) models yet in the literature. We would like to expand the partially ordered probit and logit models by changing the prior distribution, for example, to use a non-parametric or semi-parametric prior, to make the random effects $\boldsymbol{\tau}$ follow a distribution that is a realization of a Dirichlet process. Furthermore, a strategy for including prior information should be developed to help practitioners convert their "knowledge" into hyperparameters; for information on this see (Bedrick et al., 1996) and others.

In Section 4.2.3 and 5.2.3, we did multicollinearity diagnostics for the predictor variables of the models. We would like to work on developing diagnostics for detecting multicollinearity in Bayesian models for the discrete data. Current diagnostics are based on a linear model with respect to the covariates.

For the purpose of using the developed model at the University in the future, well-documented code and a "how-to" guide will be included. At a minimum this "package" will include a function to anonymize records drawn from the University (with a description of how the records were drawn), a function to read the anonymized data, and functions to fit the data (with error handling), provide descriptive summaries, descriptive diagnostics, and basic well-labeled plots. The "how-to" guide will provide how to use the functions, with a worked out example, and enough detail that a masters-level graduate student should be able to implement the model. Some detail will be provided for how to address common problems (non-convergence, etc.).

SPSS is a point-and-click spreadsheet based software which anyone with knowledge of Excel can use. To make our models easier to use by other, we will work on incorporating our new statistical models into SPSS for the Department of Mathematics and Statistics at the University to use for further evaluating of the performance of the Redesigned course. The advantage of coding the model into SPSS is that someone with only a little bit of statistical knowledge can apply the models to future Intermediate Algebra (or any other course) grade data. Currently, the model is fit in R and it would take an R novice a fair amount of time

to learn how to run the code properly. We envision that the Statistical Consulting Clinic at the University will be able to use this model to evaluate future course redesigns as well as monitoring student performance. Of course, the models with SPSS can also be easily used by other universities or colleges to analyze their course redesign or any kinds of course reform, as well as apply on any partially ordered response data like the jail time and career preference examples we mentioned earlier.

References

ALEKS home, 2013. URL `http://www.aleks.com/`.

Raymond J. Adams, Mark Wilson, and Margaret Wu. Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22(1):47–76, 1997.

Murray Aitkin and Nicholas Longford. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, pages 1–43, 1986.

James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

John Herbert Aldrich and Forrest D. Nelson. *Linear probability, logit, and probit models*, volume 45. Sage, 1984.

Paul Allison. When can you safely ignore multicollinearity. *Statistical Horizons, September*, 10, 2012.

Dorothy A. Anderson and Murray Aitkin. Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 203–210, 1985.

A.P. Basu. Identifiability. *Encyclopedia of Statistical Sciences*, 4:2–6, 1983.

Thomas Bayes. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 1763.

Brian Beaudrie. Redesigning college algebra delivery from direct instruction to a computer environment. In *Society for Information Technology & Teacher Education International Conference*, volume 2002, pages 1045–1046, 2002.

Edward J. Bedrick, Ronald Christensen, and Wesley Johnson. A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, 91(436): 1450–1460, 1996.

David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons, 2005.

James O. Berger and José M. Bernardo. On the development of reference priors. *Bayesian Statistics*, 4(4):35–60, 1992.

A.R. Bishop. *The Effect of a Math Emporium Course Redesign in Developmental and Introductory Mathematics Courses on Student Achievement and Students' Attitudes Toward Mathematics at a Two-Year College.* ERIC, 2010.

Vani K. Borooah. *Logit and probit: Ordered and multinomial models*. Number 138. Sage, 2002.

Roel Bosker and Tom Snijders. Multilevel analysis: an introduction to basic and advanced multilevel modeling. *New York*, 1999.

Eric T. Bradlow, Howard Wainer, and Xiaohui Wang. A Bayesian random effects model for testlets. *Psychometrika*, 64(2):153–168, 1999.

Norman E. Breslow and David G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

Norman E. Breslow and Xihong Lin. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):81–91, 1995.

Mark Andrew Brown. *The effectiveness of redesigning college algebra with a heavy focus on instructional technology*. PhD thesis, University of Missouri–Kansas City, 2012.

George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

Malika Chikhi. Multinomial logit models. In *THE ABSTRACT BOOK*, page 135, 2013.

Ronald Christensen. *Plane answers to complex questions: the theory of linear models*. Springer, 2011.

C. Corcoran, B. Coull, and A. Patel. EGRET for Windows user manual. *Cytel Software Corporation, Cambridge, MA*, 1999.

Stata Corporation. *Stata Statistical Software Release 7.0: Programming*. Stata Corporation, 2001.

M.K. Cowles. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6(2):101 – 111, 1996.

D. R. Cox. *The Analysis of Binary Data*. London: Methuen, 1971.

Harald Cramer. Mathematical models of statistics. *Princeton Math. Series*, (9), 1946.

Paul Damien and Stephen G. Walker. Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, 10(2), 2001.

A. Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.

Michelle R. Dedeo. Improving pass rates in mathematics using interactive computer software. In *Selected Papers from the Twelfth National Conference on College Teaching and Learning*. Center for the Advancement of Teaching and Learning, Florida Community College at Jacksonville, FL, 2001.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977.

Arthur P. Dempster, Donald B. Rubin, and Robert K Tsutakawa. Estimation in covariance components models. *Journal of the American Statistical Association*, 76(374), 1981.

Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265. ACM, 1986.

David Draper. Bayesian multilevel analysis and MCMC. In *Handbook of multilevel analysis*, pages 77–139. Springer, 2008.

Bradley Efron and Carl Morris. Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, 59(2):335–347, 1972.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitorsan empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

Bradley Efron and Carl Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Ludwig Fahrmeir and Stefan Lang. Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220, 2001.

Alan M. Ferrenberg and Robert H. Swendsen. New Monte Carlo technique for studying phase transitions. *Physical review letters*, 61(23):2635, 1988.

Antony Fielding, Min Yang, and Harvey Goldstein. Multilevel ordinal models for examination grades. *Statistical Modelling*, 3(2):127–153, 2003.

R. A. Fisher. On the mathematical foundation of theoretical statistics. Reprinted together with other papers in R.A. Fisher (1950) contribution to mathematical statistics, 1922.

R. A. Fisher. *Statistical Methods for Research Workers*. New York: Hafner, 1948.

Jean-Louis Foulley. A completion simulator for the two-sided truncated normal distribution. *Genetics Selection Evolution*, 32(6), 2000.

Andrew S. Fullerton. A conceptual framework for ordered logistic regression models. *Sociological methods and research*, 38(2):306–347, 2009.

David P. Gardner, Yvonne W. Larsen, W. Baker, et al. A nation at risk: The imperative for educational reform. *Washington, DC: US Government Printing Office*, 1983.

Alan E. Gelfand. Gibbs sampling. *Journal of the American Statistical Association*, 95(452): 1300–1304, 2000.

Alan E. Gelfand and Sujit K. Sahu. Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445): 247–253, 1999.

Alan E. Gelfand and Adrian F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

Alan E. Gelfand, Adrian F. M. Smith, and Tai-Ming Lee. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87(418):523–532, 1992.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 6:721–741, 1984.

Alexander Genkin, David D. Lewis, and David Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, 1989.

W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348, 1992.

A.R. Gilmour, R.D. Anderson, and A.L. Rae. The analysis of binomial data by a generalized linear mixed model. *Biometrika*, 72(3):593–599, 1985.

Geof H. Givens and Jennifer A. Hoeting. *Computational statistics*. John Wiley & Sons, 2012.

Peter W. Glynn and Donald L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.

Harvey Goldstein. Hierarchical data modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20(2):201–204, 1995.

Harvey Goldstein. Restricted unbiased iterative generalized least-squares estimation. *Biometrika*, 76(3):622–623, 1989.

Harvey Goldstein and Jon Rasbash. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):505–513, 1996.

P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):149–192, 1984.

William H. Greene. *LIMDEP: Version 8.0: Econometric Modeling Guide*. Econometric Software, 2002.

Hill R. Carter Griffiths, William E. and Peter J. Pope. Small sample properties of probit model estimators. *Journal of the American Statistical Association*, 82(399):929–937, 1987.

P. Haan and A. Uhlendorff. Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood. *STATA Journal*, 6(2):229 – 245, 2006.

Gary Hagerty, Stanley Smith, and Danielle Goodwin. Redesigning college algebra: Combining educational theory and web-based learning to improve student attitudes and performance. *Primus*, 20(5):418–437, 2010.

Herman O. Hartley and J. Rao. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2):93–108, 1967.

W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Donald Hedeker. Multilevel models for ordinal and nominal variables. In *Handbook of multilevel analysis*, pages 237–274. Springer, 2008.

Donald Hedeker and Robert D. Gibbons. MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer methods and programs in biomedicine*, 49(2):157–176, 1996.

Robert C. Heterick. A stone soup. *Educom Review*, 1994.

Paul W. Holland and Roy E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, 6(9):813–827, 1977.

Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

William C. Horrace and Ronald L. Oaxaca. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321 – 327, 2006.

Guan-Hua Huang and Karen Bandeen-Roche. Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32, 2004.

SAS institute. SAS/STATA® software: version 9. *Cary, NC*, 2004.

Edward H. Ip, Qiang Zhang, W. Jack Rejeski, Tamara B. Harris, and Stephen Kritchevsky. Partially ordered mixed hidden Markov model for the disablement process of older adults. *Journal of the American Statistical Association*, 2013.

Tommi S. Jaakkola and Michael I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

Jiming Jiang. REML estimation: Asymptotic behavior and related topics. *The Annals of Statistics*, 24(1):255–286, 1996.

Jiming Jiang. A derivation of BLUP: Best linear unbiased predictor. *Statistics & Probability Letters*, 32(3):321–324, 1997.

Jiming Jiang. *Linear and generalized linear mixed models and their applications*. Springer, 2007.

David Kaplan and Pamela R. Elliott. A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics*, 22(3):323–347, 1997.

Samuel Karlin and Howard M. Taylor. *A second course in stochastic processes*, volume 2. New York : Academic Press, 1981.

Ken P. Kleinman and Joseph G. Ibrahim. A semiparametric Bayesian approach to the random effects model. *Biometrics*, 54(3):921–938, 1998.

Jonathan Kropko. *Choosing between multinomial logit and multinomial probit models for analysis of unordered choice data*. ProQuest, 2007.

Pierre Simon Laplace. Memoir on the probability of the causes of events. *Statistical Science*, 1(3):364–378, 1813.

Xihong Lin and Norman E Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91 (435):1007–1016, 1996.

Ramon C. Littell. *SAS*. Wiley Online Library, 2006.

Marlaine E. Lockheed, Nicholas T. Longford, and Banco Mundial. *A multilevel model of school effectiveness in a developing country*, volume 69. World Bank, 1989.

Nicholas T. Longford. VARCL manual. *Princeton, NJ*, 1988.

Nicholas T. Longford. *Random coefficient models*. Clarendon Press, 1993.

J.S. Maritz. *Empirical Bayes methods*. Methuen, London, 1970.

Tamra Mason, Martha Byrne, Monika Nitsche, Kristin Umland, Mark Ondrias, and Karen Gardner. Proposal for Redesigning Math 120 at UNM. Technical report, The University of New Mexico, Abuquerque, NM, May 2012.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

Charles E. McCulloch. *Generalized linear mixed models*. Wiley Online Library, 2006.

Robert E. McCulloch, Nicholas G. Polson, and Peter E. Rossi. A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99 (1):173–193, 2000.

Richard D. McKelvey and William Zavoina. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4(1):103–120, 1975.

Nicholas Metropolis and Stanislaw Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

Michel Meulders, Edward H. Ip, and Paul Boeck. Latent variable models for partially ordered responses and trajectory analysis of anger-related feelings. *British Journal of Mathematical and Statistical Psychology*, 58(1):117–143, 2005.

David E. Morton, Alfred J. Saah, Stanley L. Silberg, Willis L. Owens, Mark A. Robert, and Marylou D. Saah. Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology*, 115(4):549–555, 1982.

Balgobin Nandram and Ming-Hui Chen. Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation*, 54(1-3):129–144, 1996.

J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 32:1–32, 1948.

Sean M O'Brien and David B Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746, 2004.

Ari Pakman and Liam Paninski. Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians. *Journal of Computational and Graphical Statistics*, (6), 2013.

National Mathematics Advisory Panel. *Foundations for success: The final report of the National Mathematics Advisory Panel*. U.S. Department of Education, 2008.

H. Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Output analysis and diagnostics for MCMC. *R package version 0.10-3*, 2005.

Nicholas G. Polson. *Convergence of Markov chain Monte Carlo algorithms*. University of Chicago, Graduate School of Business, Department of Economics, 1993.

Yinsheng Qu, Ming Tan, and Michael H. Kutner. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52(3):797–810, 1996.

Sophia Rabe-Hesketh, Anders Skrondal, and Andrew Pickles. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21, 2002.

Sophia Rabe-Hesketh, Anders Skrondal, and Andrew Pickles. Gllamm manual. 2004.

Sophia Rabe-Hesketh, Anders Skrondal, and Andrew Pickles. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128(2):301–323, 2005.

Adrian E. Raftery and Steven Lewis. How many iterations in the Gibbs sampler. *Bayesian Statistics*, 4(2):763–773, 1992.

Carla Rampichini and Silvana Schifini d'Andrea. A hierarchical ordinal probit model for the analysis of life satisfaction in Italy. *Social Indicators Research*, 44(1):41–69, 1998.

Jon Rasbash, William Browne, Harvey Goldstein, Min Yang, Ian Plewis, Michael Healy, Geoff Woodhouse, David Draper, Ian Langford, and Toby Lewis. *A user's guide to MLwiN*. University of London, Institute of Education, Centre for Multilevel Modelling, 2000.

Stephen W. Raudenbush, Meng-Li Yang, and Matheos Yosef. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of computational and Graphical Statistics*, 9(1):141–157, 2000.

Stephen W. Raudenbush, Anthony S. Bryk, and R.T. Congdon. Hierarchical linear modeling. *Thousands Oaks: Sage*, 2002.

Stephen W. Raudenbush, Anthony S. Bryk, and Richard Congdon. *HLM 6 for Windows*. 2004.

S.W. Raudenbush. Educational applications of hierarchical linear models: A review. *Journal of Educational and Behavioral Statistics*, 13(2):85–116, 1988.

Herbert Robbins. The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35(1):1, 1964.

Christian P. Robert. Simulation of truncated normal variables. *Statistics and computing*, 5 (2):121–125, 1995.

Christian P. Robert and George Casella. *Monte Carlo statistical methods*, volume 319. Citeseer, 2004.

Gareth O. Roberts and Sujit K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317, 1997.

German Rodriguez and Noreen Goldman. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(1):73–89, 1995.

Paul R. Rosenbaum. Some poset statistics. *The Annals of Statistics*, 19(2):1091–1097, 1991.

J. Schaafsma and David Osoba. The Karnofsky performance status scale re-examined: A cross-validation with the EORTC-C30. *Quality of Life Research*, 3(6):413–424, 1994.

Judith D. Singer. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23(4): 323–355, 1998.

D. Small. An urgent call to improve traditional college algebra programs. *MAA Online*, 2002.

Ying So and Warren F. Kuhfeld. Multinomial logit models. In *SUGI 20 Conference Proceedings*, 1995.

D.A. Sorensen, S. Andersen, D. Gianola, and I. Korsgaard. Bayesian inference in threshold models using Gibbs sampling. *Genetics Selection Evolution*, 27(3):229–249, 1995.

Larry J. Stephens and John Konvalina. The use of computer algebra software in teaching intermediate and college algebra. *International Journal of Mathematical Education in Science and Technology*, 30(4):483–488, 1999.

Gilbert W. Stewart. Collinearity and least squares regression. *Statistical Science*, 2(1):68–84, 1987.

Robert Stiratelli, Nan Laird, and James H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40(4):961–971, 1984.

Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

R. Thompson and R.J. Baker. Composite link functions in generalized linear models. *Applied Statistics*, pages 125–131, 1981.

W.A. Thompson, JR. The problem of negative estimates of variance components. *The Annals of Mathematical Statistics*, 33:273–289, 1962.

C.A. Twigg. The math emporium: A silver bullet for higher education. *Change: The Magazine of Higher Learning*, 43(3):25–34, 2011.

UNM. Official enrollment report, 2012. URL `http://registrar.unm.edu/reports--statistics/oer_s12.pdf`.

UNM. Official enrollment report, 2013. URL `http://registrar.unm.edu/reports--statistics/oer_s13.pdf`.

Paul F Velleman and Roy E Welsch. Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242, 1981.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

Christopher Winship and Robert D. Mare. Regression models with ordinal variables. *American Sociological Review*, 49(4):512–525, 1984.

George Y. Wong and William M. Mason. The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80(391):513–524, 1985.

C.F. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11 (1):95–103, 1983.

Robert G. Wynegar and Mark J. Fenster. Evaluation of alternative delivery systems on academic performance in college algebra. *College Student Journal*, 43(1):170–174, 2009.

Huiping Xu and Bruce A. Craig. A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*, 65(4):1145–1155, 2009.

Zuobing Xu and Ram Akella. A Bayesian logistic regression model for active relevance feedback. In *Proceedings of The 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 227–234. ACM, 2008.

Yuanyuan Yang and Jianchao Wang. On blocking probability of multicast networks. *Communications, IEEE Transactions on*, 46(7):957–968, 1998.

Scott L. Zeger and M. Rezaul Karim. Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American statistical association*, 86(413):79–86, 1991.

Qiang Zhang and Edward Haksing Ip. Generalized linear model for partially ordered data. *Statistics in Medicine*, 31(1):56–68, 2012.

Grietje Zuur, Paul H. Garthwaite, and Rob J. Fryer. Practical use of MCMC methods: Lessons from a case study. *Biometrical Journal*, 44(4):433–455, 2002.