STAT 778 / EDRM 828 Exam 1

Due: 4:30 pm Friday March 8th

- You may not consult or receive help from any other students or faculty on the take home exam.
- You may ask me for clarification on the questions or for some help with the computers.
- Include copies of any computer code or command files that you used, or show work done by hand. If you are using a menu driven program say which program and options you used.
- All files referred to are in the web-directory http://www.stat.sc.edu/~habing/courses/778ex1/

Note: Questions 1-3 use the data from the June 1992 Analytical Reasoning Subtest of the Law School Admissions test. The subtest had 24 questions based around four reading passages. Questions 1-6 were based on the first passage, questions 7-11 were based on the second passage, questions 12-17 were based on the third passage, and questions 18-24 were based on the final passage. A sample of 1,000 examinees' responses to this test can be found in two different formats in the files `a692a.dat` and `a692b.dat` .

Question 1: Use either α or λ_2 to give an estimate of the reliability of this test.

a) Briefly say why you chose the statistic (α or λ_2) you did
b) Report your estimate of the test's reliability
c) Comment on any weaknesses of using your choice of statistic as an estimate of reliability. (You do not need to comment on any weaknesses in the concept of reliability itself, just on the estimate of reliability.)

Queston 2: Another method of estimating reliability is to take the correlation of the examinees' observed total scores on a test with their observed total scores on a parallel test. We could perform this estimation by splitting the 24-item exam into two 12-item parts that are felt to be reasonably parallel; finding the correlation of the observed scores of the examinees on the two 12-item parts; and then adjusting that figure back to what it should be for a 24-item test.

a) Use either classical or IRT methods to determine two groups of twelve items that you feel should lead to tests that are reasonably parallel. Briefly justify the way you chose the two groups. (Code is on the web site for actually splitting the test into those groups.)
b) Use the observed total scores on these two 12-item tests to estimate the 12-item test reliability
c) Use the estimate of the 12-item test reliability to estimate the reliability of the original 24-item test

Question 3: A final way to estimate the reliability of a test is to find the square of the correlation coefficient between the observed total test score and the examinees true score.

In reality we can't do this of course because we don't know the examinees true scores! We could however use IRT to estimate the IRFs and examinee abilities, and then use those estimates, along the assumptions of 3PL form and local independence, to simulate a new test. (Such a simulated test is saved as `a692sim.dat` .)

Because we simulated the test we could calculate the true score and would also have the observed score.

a) Based on the description of the real test and the way the simulated test was generated, in what way do the two tests differ? Why would this lead one to expect that the simulated test would have a higher reliability?

b) In spite of the answer to part a, if you actually redo question 1a on `a692sim.dat` set you get a lower reliability for the simulated test! What characteristic of the examinees changed between the real data set and simulated data set to cause this? Why was this characteristic different in the simulated data set than in the real data set?

Question 4: Consider a two-item test, where both items follow the 2PL model. The first item has parameters a=1 and b=0.5, and the second has parameters a=1.5 and b=-0.5.

a) Determine the probability that examinees of ability -2, -1, 0, 1, and 2 would correctly answer each item
b) Determine the true score on this two-item test for examinees with ability -2, -1, 0, 1, and 2
c) Determine the probability that examinees of ability -2, -1, 0, 1, and 2 would have observed scores patterns of (0,0), (1,0), (0,1), (1,1) respectively
d) Using only the your answer to part c, give a crude approximation of the MLE estimate of ability for examinees who had response patterns of (0,0), (1,0), (0,1), and (1,1) respectively.

Question 5: The data sets `simrasch.dat` and `sim3pl.dat` each contain the simulated data from 1000 examinees taking a ten-item test. The parameters used come from Appendix A in the text, and were estimated from a New Mexico State Proficiency Exam. The examinee abilities used in the simulation study are stored in `simab`.

	RASCH **(a=0.59)**		3PL	
	b	a	b	c
1	-1.43	0.64	-1.10	0.17
2	-0.99	0.94	-0.50	0.17
3	-0.51	0.93	-0.09	0.17
4	-0.50	0.69	-0.09	0.17
5	0.12	0.41	0.44	0.17
6	0.17	1.23	0.76	0.25
7	0.33	1.50	0.58	0.17
8	0.33	0.99	0.68	0.17
9	1.11	0.63	1.54	0.10
10	1.24	1.14	1.46	0.15

a) Fit both the Rasch and 3PL model to the `simrasch.dat` data set using BILOG. Compare how well each of these two models is able to recover the examinee abilities when the data actually follows the Rasch model. (Be sure to say how you were measuring these accuracies.)

b) Fit both the Rasch and 3PL model to the `sim3pl.dat` data set using BILOG. Compare how well each of these two models is able to recover the examinee abilities when the data actually follows the 3PL model. (Be sure to say how you were measuring these accuracies.)

c) Based on your results in parts a and b, what model would you recommend to someone who was giving a short exam and wasn't sure if guessing was a factor or not? Why?