STAT-S632 – Midterm Exam – Take-home

Due April 30, 2018

Instructions:

Please read carefully the following instructions:

- Start your document with the statement: "With the exception of the class instructor, I have not had any form of communication about this exam with any other individual (including other students, teaching assistants, instructors, etc.)" and sign by hand or typing your name.
- You will be allowed to ask the instructor clarification questions about this exam only during class sessions. This ensures that all students receive the same information about the exam. No questions about this exam will be answered during individual office hours or additional review sessions.
- Your are allowed to use course resources without need of citations. The use of additional resources (e.g., papers, manuscripts, textbooks, internet, etc.) needs to be cited appropriately. Failure to cite resources could be penalized.
- Your document should be prepared in the following way:
  - Final answers should be included in the **first 2 pages** of your submission paper.
  - Supporting material (R syntax used or additional output, graphs, etc.) is also required and should be added after the first two pages. In addition, if any given output was used to answer questions, the answer should include a clear reference to the material location (e.g., figure number, table number, or at least page number). There is no limit of space for supporting material.
- Failure to follow any of the instructions above could result in reduction of your grades.
- Submit your solutions to Canvas as a single PDF file by midnight on Monday, April 30th.

Questions

The file `S18S632final.txt` contains data on 4106 grade-8 students (who are approximately 11 years old) in 216 primary schools in the Netherlands. The data set includes the following variables:

- `school`: a (non-consecutive) ID number indicating which school the student attends.
- `iq`: the student's verbal IQ score, ranging from 4 to 18.5 (i.e., not traditionally scaled to a population mean of 100 and standard deviation of 15).
- `test`: the student's score on an end-of-year language test, with scores ranging from 8 to 58.
- `ses`: the socioeconomic status of the student's family, with scores ranging from 10 to 50.
- `class.size`: the number of students in the student's class, ranging from 10 to 42; this variable is constant within schools, apparently reflecting the fact that all of the students in each school were in the same class.
- `meanses`: the mean SES in the student's school, calculated from the data; the original data set included the school-mean SES, but this differed from the values that I computed directly from the data, possibly it was based on all of the students in the school.
- `meaniq`: the mean IQ in the student's school, calculated (for the same reason) from the data.

There are some missing data, and I suggest that you begin by removing cases with missing data. Then add the following two variables to the data set:

- school-centered SES, computed as the difference between each student's SES and the mean of his or her school; and
- school-centered IQ.

1. Examine scatterplots of students' test scores by centered SES and centered IQ for each of 10 randomly sampled schools. Do the relationships in the scatterplots seem reasonable linear? Hint: In interpreting these scatterplots, take into account the small number of students in each school, ranging from 4 to 34 in the full data set.
2. Regress the students' test scores on centered SES and centered IQ within schools for the full data set - that is, compute a separate regression for each school. Then plot each set of coefficients (starting with the intercepts) against the schools' mean SES, mean IQ, and class size. Do the coefficients appear to vary systematically by the schools' centered SES, centered IQ, and class size)?
3. Fit linear mixed-effects models to the data, proceeding as follows:
   a. Begin with a model of test scores by schools with no fixed effects and only a random intercept. What proportion of the total variation in test scores among students is between schools (i.e., what is the intraclass correlation)?
   b. Obtain a model for regressing test scores by schools with any fixed effects at the student level you consider appropriate and random effects for students' centered SES and centered IQ (recall that your only grouping variable is school). Test whether each of these random effects is needed, and eliminate from the model those that are not (if

any are not). How, if at all, are test scores related to the predictors? Justify your choices/decisions. Note: You may obtain a convergence warning in fitting one or more of the null models that remove variance and covariance components; this warning should not prevent you from performing the likelihood-ratio test for the corresponding random effects.

c. Introduce school level predictors; i.e., mean school SES, mean school IQ, and class size in your model. Test whether the random effects that you retained in part b model are still required now that there are other predictors in the model. Note: Again, you may obtain a convergence warning.
d. Compute tests of the various main effects and interactions. Then simplify the model by removing any fixed-effects terms that are nonsignificant.
e. Interpret the results obtained for the simplified model.