Eco311 Optional Reading: Catch me if you can

(Jing Li, Miami University)

1. The econometric methods we learn in eco311 can be used to prove causality, or something else. In this note we focus on using statistics to reveal or detect pattern in data, and more specifically, unusual pattern (data inconsistency) in the answer sheet of an exam, which may indicate cheating behaviors of either students or teachers. You may read Chapter 1, page 22–35, of *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything* written by Steven D. Levitt and Stephen J Dubner for a fascinating story of catching TEACHERS who cheat in Chicago.

2. Suppose there are ten multiple choice questions in an exam. The correct answers are

   $$ACDAAB \quad CDCA$$

   The answers of one student are

   $$BCCABB \quad CDCA$$

   If I tell you the first six questions are easy, while the last four questions are hard. Can you see something strange in the student's answers? To highlight the pattern, let's transform the data to a dummy variable—0 for answer being wrong and 1 for answer being correct:

   $$010101WOW1111$$

   I put a WOW in the middle to divide the sample into two sub-samples—easy one before WOW and difficult one after WOW. Can you see the unexpected pattern now?

3. This student is UNBELIEVABLE by having 100 percent right answers for difficult problems, but only 50 percent for easy ones. Interesting! Let me use the jargon you hate. In this case we are conducting a two sample t test by comparing the mean (proportion) across two sub-samples.

4. To investigate this matter further, I also look at the answer of the student sitting next to the first student. His answer is

   $$ACBAAB \quad CDCB$$

Again, in order to highlight the pattern, let's transform the data first. This time I want to convert letters into numbers—1 for A, 2 for B, 3 for C, and 4 for D. The numeric answers for the two students are

233122 3431

132112 3432

What catches my eye is that for the easy questions, the answers of two students are less correlated than the difficult questions. I can use jargon again: You may run two (group-wise) regressions—one for easy sub-sample and the other for difficult sub-sample. More explicitly, first you regress (233122) onto (132112). Then you regress (3431) onto (3432). The R-squared for the second regression is greater than the first regression. Interesting !!

5. Exercise: do you think the second student is likely to be a cheater or not? Why? (Hint: consider two-sample t test applied to the second student)

6. Here comes the nuclear weapon—suppose we have a score spreadsheet for 100 students. I know how many points each student gets in exam 1, exam 2 and exam 3. That means I can predict exam 3 based on exam 1 and exam 2. The potential cheaters in exam 3 may be those who perform "much much" better than the exam 1 and exam 2 would predict. In terms of econometrics, I want to get the residual, sort the data, and scrutinize students with greatest positive residuals.

7. To summarize, here I discuss how to use three basic statistical tools to spot irregularity in data. Fancier methods can be tried, but simple ones may do an equally good job. The most important thing is thinking, other than econometrics. For this problem, you have to think what can be served as the evidence or hint for cheating. What transformation can be applied to data to highlight the pattern? What statistical tools are appropriate? How to interpret the statistical findings?