# PS C236A / Stat C239A
# Problem Set 4
# Due: Oct. 29, 2012

## Instructions

This assignment is due **4 pm Monday, Oct. 29.** You may submit your analytical work either electronically or in paper form. Electronic versions must be sent as a .pdf to <jahenderson[at]berkeley.edu>. Paper copies should be placed in my mailbox in 210 Barrows. For the computing portion of the assignment, you must submit a fully executable version of all .R code, along with any data used in the code (excepting that provided through the course webpage) to the email above. All files for each assignment sent electronically should be included in one omnibus email, with the subject line containing the course and homework number, and your last name (e.g., PS239A/STAT236A: HW4 - McConnell).

You are encouraged to work together in groups to complete the assignments. However, you must hand in your own individual answers. Photocopies and other reproductions of someone else's answers are not acceptable. Please also list the names of everyone with whom you have collaborated on this assignment.

**Problem 1** Eggers and Hainmueller (2009) estimate the LATE of just barely winning (losing) an election contest on wealth at death. Their estimate compares the average wealth at death of all winners of contests between 1950-1970 to the average wealth over all losers over that time. There were seven general elections over that period.

Consider another estimator of the LATE that takes the average difference between the winners and the losers participating in a given general election, and takes the weighted average of these differences—weights are:

$$\frac{\#\text{near-winners and near-losers in that general election}}{\text{total number of near-winners and near-losers across all elections}}.$$

Each near-win (near-loss) candidate is only included in this average once, according to the contest in which that candidate is ultimately classified as near-win or near-loss.

a. Write out mathematical expressions for both of these estimators (i.e., Eggers and Haimueller (2009) and the weighted estimator).

b. Show that, if the number of contests within each general election is the same, and the number of near winners is equal to the number of near losers in every general election, then the two estimates are the same.

c. Suppose the same assumptions as part (b), except that the 1950 general election had twice as many contests as all other elections and had twice as many near-winning candidates as near-losing candidates. In general, are the two estimators the same? Or are they different?

**Problem 2** Suppose that 1,000,100 students take an exam. Students can score an integer number of points between 0 and 10,000 on the exam: the set of possible test scores is $\{0, 1, \ldots, 9, 999, 10, 000\}$. Miraculously, for each possible point value of the exam, exactly 100 students score that many points. Those students that score above a certain threshold on the exam (usually around 5,000 points) receive a scholarship. We are interested in estimating the LATE of receiving a scholarship on future earnings, for those students that score 5,000 points.

Let $Y_i(1)$ and $Y_i(0)$ denotes the future earnings of student $i$ when that student receives (or does not receive) the scholarship. Let $s_i$ denote the test score of student $i$. Let $c_i = 1$ if student $i$ receives a scholarship and $c_i = 0$ if that student does not receive a scholarship Suppose that the distribution of $IQ$ for those students scoring 4,995 points is the same for students scoring 4,996 points, 4,997 points, . . ., 5,004 points, 5,005 points. We want to estimate:

$$\text{LATE} = E(Y_i(1) - Y_i(0)|s_i = 5,000)$$

a. Suppose that all students that score above 5,000 points receive a scholarship, and all students that score below 5,000 points do not receive a scholarship. For students that score exactly 5,000 points, half will be randomly selected to receive a scholarship, and half will not receive that scholarship. Give an unbiased estimate of the LATE. Does this estimate require smoothness of $E(Y_i(1))$ and $E(Y_i(0))$ at the $s_i = 5,000$ threshold?

b. Suppose for parts (b), (c), and (d) that all students scoring 5,000 points or above receive a scholarship.

   Suppose that future earnings are determined by the following model:

   $$Y_i = \alpha + \beta_1 s_i + \beta_2 c_i + \beta_3 s_i c_i + \beta_4 IQ_i + \epsilon_i$$

   where $\epsilon_i$ has expectation 0 and variance $\sigma^2$. Give an unbiased estimate of the LATE. Under this model, are $E(Y_i(1))$ and $E(Y_i(0))$ smooth at the $s_i = 5,000$ threshold? Are these assumptions stronger than those required for regression discontinuity?

c. Suppose that, for scores between 4,995 and $5,000 - \epsilon$ points, future earnings are determined by

   $$Y_i = 50,000 + 5,000(s_i - 5,000) + \epsilon_i$$

   and for scores between 5,000 and 5,005 points, future earnings are determined by

   $$Y_i = 80,000 - 6,000(s_i - 5,000) + \epsilon_i.$$

   Give an unbiased estimate of the LATE. In what sense are these assumptions stronger or weaker than those in (b)? Are these assumptions stronger than those required for regression discontinuity?

d. Suppose some students that would ordinarily receive low test scores cheat off of good students. All students that cheat score 5,000 points or above, with some students scoring exactly 5,000 points. Assume that, had the students not cheated, the assumptions for regression discontinuity analysis would hold. After the students cheat, do these assumptions still hold? Why or why not? What if all of the cheating students scored above 5,002 points?

**Problem 3** This question will involve the RD design controversy in Lee (2008) and Caughey and Sekhon (2011). The following file is on bspace (LeeRDdata.zip), and contains all the necessary data to answer the question.

Note that Lee's dataset is different from that used by Caughey and Sekhon (2011). For example, it includes far fewer variables and it contains some errors and missing values that were imputed. Both the Caughey and Sekhon article and an appendix that includes details about their dataset are available on Sekhon's webpage.

a. Use David Lee's replication files to replicate the tables and figures in Lee's article.

b. To the extent possible, use Lee's dataset to replicate the key tables and figures in Caughey and Sekhon (2010). Which key findings differ between Lee (2008) and Caughey and Sekhon (2010) because of data differences and which findings are consistent even if one uses Lee's original dataset?

Bonus On the same data, show whether or not previous *incumbent* win margin is smooth through the cut-point in the design using McCrary's test of smoothness outlined in his 2008 paper: `http://emlab.berkeley.edu/~jmccrary/mccrary2006_DCdensity.pdf`. Now show this for *Democratic candidate* previous vote margin.