Econ 270C: Analytics of Economic Development
Problem Set 1 (Due Tuesday April 3, 2007)

Child Health and Education: Non-parametric regressions

For this problem, download the STATA dataset "PSET1-2007.DTA" from the course page. The data are from a joint project with Michael Kremer and Rebecca Thornton in rural Kenya (on the course syllabus, "Incentives to Learn", although note that the data you will use here is only a subset of the overall dataset). The data here are from a recent follow-up survey of the "Incentives to Learn" paper sample.

This problem examines the relationship between merit awards and later academic performance among Kenyan schoolchildren and young adults. In early 2001, Grade 6 girls in a random subset of "treatment" schools (variable name "treat") were offered a large cash award, including a school fee subsidy, if they scored in the top 15% of all treatment school girls. These girls are cohort 1 (indicator variable "c1"). The same program was repeated for Grade 6 girls in 2002 (cohort 2). The scholarship winners in 2001 (2002) are denoted by the variable "winn01" ("winn02").

The dataset contains test score information from late 2000, the year before the program through late 2002. The test score outcomes ("test00", "test01", "test02") were normalized such that the test distribution in the comparison schools is mean zero with a standard deviation of one (for all students in that grade, not just those in this sample – thus the mean need not equal 0 here).

We also have information gathered in 2005-2006 in a follow-up survey, designed to measure longer term impacts of the program. These variables include a variable for whether or not the girl is still in school ("in_school") and their educational attainment ("educ_attain") at the time of the follow-up, as well as test score data ("vocab" for their normalized score on English and Swahili vocabulary tests, and "math" for an arithmetic test.).

The goal of this problem set is to understand medium-term impacts of the program – both the incentive component, and the impact of "winning" – using various parametric and non-parametric econometric methods. Another important issue for the analysis is whether girls at the bottom of the baseline test score distribution were harmed by the program – perhaps due to demoralization or diversion of teacher attention to high-achieving classmates, for instance.

Please include all regression output and graphs, as well as do-files, with your solutions.

a) Present summary statistics for all the variables in the dataset. Do baseline 2000 test scores differ on average across the treatment and comparison students? Do 2001 test scores differ on average across treatment groups? Do 2002 test scores differ on average across treatment groups? Do the two 2005-2006 test scores differ on average across treatment groups?
[1 point] (STATA command hints: "summarize, detail", "ttest")

b) Plot the kernel density of 2000 test scores in the following ways:

(i) Epanechnikov kernel with the optimal bandwidth
(ii) Epanechnikov kernel with bandwidth equal to 0.05
(iii) Gaussian kernel with optimal bandwidth

Characterize the distribution of 2000 test scores. Does kernel density estimation appear to be more sensitive to the bandwidth or to the kernel? **[1 point]**
(STATA command hints: "kdensity", "graph twoway")

c) Plot the kernel density of 2000 test scores with Epanechnikov kernel and optimal bandwidth separately for treatment and comparison students, and place them in one figure.

Create a new variable (test05) that is the average of the 2005-2006 vocabulary and math test scores. Plot the kernel density of this 2005-2006 test score with Epanechnikov kernel and optimal bandwidth separately for treatment and comparison students, and place them in one figure. Restrict attention to only those students who also have 2000 test scores. Describe any shifts in the test distributions through time. **[1 point]**

d) Determine the linear relationship between the 2005-2006 test score "test05" (as dependent variable) and the 2000 test score using OLS, and present the regression results.

Plot the predicted quadratic relationship between the 2005-2006 test scores and the 2000 test scores, as well as 95 percent confidence bounds.

Plot the predicted quadratic relationship between the 2005-2006 and 2000 test scores separately for treatment and comparison students, and place them in one figure. **[1 points]**
(STATA command hints: "regress", "graph", "qfit / qfitci")

e) Perform the following non-parametric Lowess regressions of the average 2005-2006 test score, test05, on the 2000 test score:

(i) All students, bandwidth equal to 0.5
(ii) All students, bandwidth equal to 0.05
(iii) Treatment students, bandwidth equal to 0.5
(iv) Comparison students, bandwidth equal to 0.5

Characterize the non-parametric relationship between the 2005-2006 test score and 2000 test score. Does the relationship appear to differ for treatment versus comparison students? Are there important patterns that were not apparent from the linear or quadratic fits in part (d)? **[1 point]** (STATA command hints: "lowess", "graph, twoway")

f) Perform the Fan locally-weighted non-parametric regression, using a quartic kernel with bandwidth equal to 0.5. Include 95 percent confidence bounds in all plots, bootstrapping the standard errors.[1] "Trim" extreme values from the distribution, only considering the 2000 test score interval of [-1.3, 1.6], for simplicity.

(i) Treatment pupils, regress the 2005-2006 average test score (as dependent variable) on the 2000 test.
(ii) Comparison pupils, regress the 2005-2006 average test score (as dependent variable) on the 2000 test.

[1] Refer to *The Analysis of Household Surveys* by Angus Deaton for the STATA code on Fan regressions and bootstrapping. These are also online at *www.worldbank.org/LSMS/tools/deaton*. Be aware that there are a few errors in the published Fan regression code that you will need to spot.

(iii) Take the difference between these two non-parametric relationships (for each test), and bootstrap the standard errors of this difference, for various points throughout the 2000 test score distribution.

Is the 2005-2006 test score difference between treatment and comparison students significantly different than zero (at over 95 percent confidence) anywhere in the baseline 2000 test score distribution? Please describe. **[2 points]**

g) Estimate the reduced-form impact of treatment assignment among cohort 1 (c1==1) students: regress the indicator for educational attainment by 2005-2006 on the treatment indicator using OLS (taking into account any common error component within schools). Is this relationship robust to controlling for the baseline (2000) test score? Do any subgroups of girls, by initial age ("age01"), say, gain more than others from the incentive program? Present the regression results and discuss. **[1 point]** (STATA command hints: "cluster")

h) Conduct a regression discontinuity (RD) analysis of the impact of winning the award in 2002 on educational attainment in 2005-2006. (Restrict attention to cohort 2, in the treatment schools only.) Control for fourth-order polynomial trends in the 2002 test score in both sides of the winning threshold of 0.63 s.d. on the 2002 test, and estimate the discontinuity at this point in a regression. (Winners are denoted by the indicator variable "winn02".) Also present the figure corresponding to this regression, with the appropriate 95% confidence intervals. Does winning the award appear to affect future educational attainment? (Hint: refer to the RD STATA code on Enrico Moretti's website, for his 2004 *QJE* paper, for this type of graphical RD analysis: http://www.econ.berkeley.edu/~moretti/papers.html) **[2 points]**