PS C236A / Stat C239A
Problem Set 4 - Solutions

1) Let N_i denote the number of near-winners and near-losers for the ith general election. Let W_{ij} denote the wealth at death for the jth candidate in the ith general election. Let Win_{ij} denote an indicator variable: $Win_{ij} = 1$ if the jth candidate in the ith general election is a near-winner, otherwise $W_{ij} = 0$. Let $N = \sum N_i$ denote the total number of near-winners and near-losers. Let $\#Win_i = \sum_j Win_{ij}$ denote the total number of winners in the ith general election and let $\#Win = \sum_i W_i$ denote the total number of winning candidates across all elections.

   a) The Eggers and Hainmueller estimator is

$$\sum_{i=1}^{7}\sum_{j=1}^{N_i}\left(\frac{W_{ij}Win_{ij}}{\#Win} - \frac{W_{ij}(1-Win_{ij})}{N-\#Win}\right) \tag{1}$$

   The estimator described in the problem is

$$\sum_{i=1}^{7}\frac{N_i}{N}\sum_{j=1}^{N_j}\left(\frac{W_{ij}Win_{ij}}{Win_i} - \frac{W_{ij}(1-Win_{ij})}{N_i-Win_i}\right) \tag{2}$$

   b) There is a slight typo in this problem. Instead of "number of contests," it should be "number of near-winners and near-losers."

   There are a total of $N/2$ near-winners and $N/2$ near-losers. For each election, there are a total of $N/7$ contests. The estimator (1) becomes

$$\sum_{i=1}^{7}\sum_{j=1}^{N/7}\left(\frac{W_{ij}Win_{ij}}{N/2} - \frac{W_{ij}(1-Win_{ij})}{N/2}\right)$$

   The estimator (2) becomes

$$\sum_{i=1}^{7}\frac{1}{7}\sum_{j=1}^{N/7}\left(\frac{W_{ij}Win_{ij}}{N/14} - \frac{W_{ij}(1-Win_{ij})}{N/14}\right) = \sum_{i=1}^{7}\sum_{j=1}^{N/7}\left(\frac{W_{ij}Win_{ij}}{N/2} - \frac{W_{ij}(1-Win_{ij})}{N/2}\right)$$

   Thus, the estimators are the same.

   c) The fraction of winning candidates is:

$$\frac{6}{8}\left(\frac{1}{2}\right) + \frac{2}{8}\left(\frac{2}{3}\right) = \frac{13}{24}$$

   Thus, for the estimator in (1), every winning unit receives weight $\frac{24}{13N}$. However, for the estimator in (2), winning people in 1950 have coefficient

$$\frac{2}{8}\frac{1}{\frac{2}{3}\frac{2}{8}N} = \frac{3}{2N}$$

   Thus, winning people in 1950 have weight $3/2N$. Similarly, winners in every other year have weight $2/N$. And so, the estimators are not the same.

2) a) An unbiased estimate of the LATE is

$$\sum_{i=1}^{N}\left(\frac{Y_i(1)(c_i)\mathbf{1}(s_i=5,000)}{50}-\frac{Y_i(1)(1-c_i)\mathbf{1}(s_i=5,000)}{50}\right)$$

There are no smoothness conditions necessary, as there is some people at the cutpoint $s_i = 5000$, with half of those people receiving scholarships, and the other half not receiving scholarships.

b) According to the model, and supposing equal distribution of $\mathbb{E}(IQ_i)$ across $c_i = 1$ and $c_i = 0$, the LATE is:

$$\begin{aligned}
LATE &= \mathbb{E}(Y_i(1)-Y_i(0)|s_i=5000) = \mathbb{E}(Y_i|c_i=1,s_i=5000)-\mathbb{E}(Y_i|c_i=0,s_i=5000)\\
&= \mathbb{E}(\alpha+\beta_1 s_i+\beta_2 c_i+\beta_3 s_i c_i+\beta_4 IQ_i+\epsilon_i|c_i=1,s_i=5000)\\
&\quad -\mathbb{E}(\alpha+\beta_1 s_i+\beta_2 c_i+\beta_3 s_i c_i+\beta_4 IQ_i+\epsilon_i|c_i=0,s_i=5000)\\
&= \alpha+5000\beta_1+\beta_2+5000\beta_3+\beta_4\mathbb{E}(IQ_i|c_i=1,s_i=5000)\\
&\quad \alpha+5000\beta_1+\beta_4\mathbb{E}(IQ_i|c_i=0,s_i=5000)\\
&= \beta_2+5000\beta_3
\end{aligned}$$

We can estimate the coefficients of the linear model unbiasedly using OLS. Thus, the estimate of LATE is

$$\hat{\beta}_2+5000\hat{\beta}_3$$

This model insures smoothness of $E(Y_i)$ at the cutpoint. The assumption that IQ_i is smooth is assumed by the problem description. And treatment is only assigned if $s_i \geq 5000$. Thus, all assumptions necessary for regression discontinuity analysis are met; the assumption that points follow this model is stronger than required for regression discontinuity.

c) The LATE is

$$LATE = \mathbb{E}(Y_i(1)-Y_i(0)|s_i=5000) = 80,000-50,000 = 30,000$$

If we did not know this equation, we could estimate both lines using OLS, and take the difference of the intercept terms. These assumptions are weaker than those in part b), only because we are assuming that the model holds for test scores between 4,995 and 5,005 as opposed to holding over all test scores, and does not depend on IQ; otherwise, the assumptions are exactly the same. The linear model for test scores over the entire range is

$$Y_i = 50000+30000c_i+5000(s_i-5000)-11000(s_i-5000)c_i$$

The fact that the slopes are dramatically different before and after the cutpoint does not matter, this is simply saying that there is a negative score by treatment interaction effect. By the same arguments as before, these assumptions are stronger than those for regression discontinuity.

d) By low test scores, it is implied that these are scores that are far below 5,000 points. The assumptions of covariate smoothness and continuity of $\mathbb{E}(Y_i)$ are both violated. The covariate "Test score before cheating," which likely effects future income, is guaranteed to be imbalanced at the cutpoint, since everyone before the cutpoint will have a test score before cheating equal to their test score, and some people after the cutpoint will have a test score before cheating far lower than their true test-score. Also, if future income is positively correlated with test score before cheating, then future outcome is likely to have non-continuous jumps after the cutpoint, including at the cutpoint.

If everyone who cheats scores 5002 points or above, there is no violation of RD assumptions; all violations of continuity and covariate balance occur away from the cutpoint. However, there is much less data that can be used for model fitting; fitting a model within the ± 5 test-score window may introduce bias in an estimate of LATE.

3) See `HW4_Answers.R` for solutions