Regression Discontinuity Simulation Homework

Jennifer Hill, Ray Lu & Zarni Htet

Objective

The goal of this exercise is to simulate and analyze data that might have arisen from a policy where eligibility was determined based on one observed measure. Data from this type of setting are often consistent with the assumptions of the regression discontinuity designs we discussed in class.

Setting

This assignment simulates hypothetical data collected on women who gave birth at any one of several hospitals in disadvantaged neighborhoods in New York City in 2010. We are envisioning a government policy that makes available pre- and post-natal (through 2 years post-birth) health care for pregnant women, new mothers and their children. This program is only available for women in households with income below $20,000 at the time they gave birth. The general question of interest is whether this program increases a measure of child health at age 3. You will generate data for a sample of 1000 individuals.

Clean regression discontinuity design. For this assignment we will make the unrealistic assumption that everyone who is eligible for the program participates and no one participates who is not eligible.

Question 1. God role: simulate income.

Simulate the "assignment variable" (sometimes referred to as the "running variable", "forcing variable", or "rating"), income, in units of thousands of dollars. Call the variable "income". Try to create a distribution that mimics the key features of the data displayed in income_hist.pdf.

Question 2. Policy maker role: Assign eligibility indicator.

Create an indicator for program eligibility for this sample. Call this variable "eligible".

Question 3: God role.

For question 3 you will simulate a health measure with a minimum possible score on *observed data* of 0 and maximum possible score of *30*. You will simulate data from two possible worlds that vary with regard to the relationships between health and income.

Question 3a

(a) God role. Simulate potential outcomes for World A.

i) Generate the potential outcomes for health assuming linear models for both E[Y(0) | X] and E[Y(1) | X]. This health measure should have a minimum possible score of 0 and maximum possible score of 30. The *expected* treatment effect for everyone should be 4 (in other words, E[Y(1) - Y(0) | X] should be 4 at all levels of X). The residual standard deviation of each potential outcome should be *2*.

ii) Save two datasets: (1) fullA should have the forcing variable and both potential outcomes and (2) obsA should have the forcing variable, the eligibility variable, and the observed outcome.

Question 3b

(b) Simulate potential outcomes for World B.

i) Generate the potential outcomes for health assuming a linear model for E[Y(0) | X] and a quadratic model for E[Y(1) | X]. The treatment effect at the threshold (the level of X that determines eligibility) should be 4. The residual standard deviation of each potential outcome should be *2*. Creating this DGP may be facilitated by using a transformed version of your income variable that subtracts out the threshold value.

ii) Save two datasets: (1) fullB should have the forcing variable and both potential outcomes and (2) obsB should have the forcing variable, the eligibility variable, and the observed outcome.

Question 4. Researcher role. Plot your data!

Make two scatter plots of income (x-axis) versus observed health (y-axis), one corresponding to each world. In each, plot eligible participants in red and non-eligible participants in blue.

Question 5. Researcher role. Estimate the treatment effect for World A and World B using all the data.

Now we will estimate effects in a number of different ways. Each model should include reported income and eligible as predictors. In each case use the model fit to report the estimate of the effect of the program at the threshold level of income. All models in Question 5 will be fit to all the data.

Question 5a: Researcher role. Estimates for World A using all the data.

(a) Using all the data from World A, perform the following analyses.

(b) Fit a linear model to the full dataset. Do not include an interaction.

(ii) Fit a linear model to the full dataset, include an interaction between income and eligible.

(iii) Fit a model that is quadratic in income and includes an interaction between both income terms and eligible (that is – allow the shape of the relationship to vary between treatment and control groups.

Question 5b: Researcher role. Estimates for World B using all the data.

(b) Using all the data from World B, perform the following analyses.

(c) Fit a linear model to the full dataset. Do not include an interaction.

(ii) Fit a linear model to the full dataset, include an interaction between income and eligible.

(iii) Fit a model that is quadratic in income and includes an interaction between both income terms and eligible (that is – allow the shape of the relationship to vary between the treatment and control groups).

Question 6. Researcher role. Estimate the treatment effect for World A and World B using data close to the threshold.

We will again estimate effects in a number of different ways. Each model should include "income" and "eligible" as predictors. In each case use the model fit to report the estimate of the effect of the program at the threshold level of income. All models in Question 6 will be fit only to women with incomes ranging fro $18,000 to $22,000.

Question 6a: Researcher role. Estimates for World A using the restricted data.

(a) Using the restricted data (for participants with incomes between $18K and $22K) from World A, perform the following analyses.
(b) Fit a linear model to the restricted dataset. Do not include an interaction.

(ii) Fit a linear model to the restricted dataset, include an interaction between income and eligible.
(iii) Fit a model that is quadratic in income and includes an interaction between both income terms and eligible (that is – allow the shape of the relationship to vary between the treatment and control groups).

Question 6b: Researcher role. Estimates for World B using the restricted data.

(b) Using the restricted data (for participants with incomes between $18K and $22K) from World B, perform the following analyses.

(c) Fit a linear model to the restricted dataset. Do not include an interaction.

(ii) Fit a linear model to the restricted dataset, include an interaction between income and eligible.
(iii) Fit a model that is quadratic in income and includes an interaction between both income terms and eligible (that is – allow the shape of the relationship to vary between treatment and control groups.

Question 7. Researcher role. Displaying your estimates.

Present your estimates from questions 5 and 6 into one or two tables or figures, clearly noting which world the data are from, which models the estimates are from, and which analysis sample was used.

Question 8. Researcher role. Thinking about the data.

(a) A colleague now points out to you that some women may have incentives in these settings to misreport their actual income. Plot a histogram of reported income (using the default settings which should give you 33 bins) and look for anything that might support such a claim. What assumption is called into question if women are truly misreporting in this manner (choose the best answer below)?

(b) Another colleague points out to you that several other government programs (including food stamps and Headstart) have the same income threshold for eligibility. How might this knowledge impact your interpretation of your results?

Question 9. Researcher role. Thinking about the assumptions?

What are the three most important assumptions for the causal estimates in questions 5 and 6?

Question 10.

Provide a causal interpretation of your estimate in Question 6biii.

Challenge Problem.

Use the rdrobust package in R to choose an optimal bandwidth for the data in World B (can use any of the approaches they support). Two points for each one you try for a maximum of 6 points.