AEB 6933 problem set 2

In this problem set we will replicate and extend results from the following paper:

Kremer, M. and E. Miguel. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica*, 72(1): 159-217.

You are welcome to work together. Please turn in your code and any graphs or tables produced while completing the assignment. If you work with a partner, be sure both of your names are on the code. The code should organize your output by question number so I can easily follow your results. When I ask you for a written response, you may enter these directly into your code as text.

1. We will begin by estimating differences in means between schools assigned to treatment in 1998 (Group 1) and schools assigned to treatment in 1999 (Group 2), prior to receipt of treatment by Group 2 schools. We will use *any_ics99*, which is equal to 1 for all students with any moderate to heavy helminth infections in 1999, as our dependent variable.

   Open the data set *ps2.dta*. Estimate the difference in means for *any_ics99* using a weighted linear regression. Use *indiv_weight* as your weighting variable. This variable ensures that our sample is representative of the population of students from which our sample is drawn. Cluster your standard errors by *sch98v1* (we will discuss clustering in detail later, the short answer is we are letting the regression error be correlated within groups defined by *sch98v1*, i.e., school). I used feols from the "fixest" package and specified "vcov = ~ sch98v1".

2. A common way to perform robustness checks on a causal effect estimate is to check sensitivity to mode of inference, i.e., how you perform your hypothesis tests. In this question, we will use randomization inference to test the difference in means described in the previous question. That is, we will re-randomize treatment many times, and each time through estimate the difference in means described in question 1. We will save the resulting estimates and use them to obtain the distribution of the estimated difference in means under the "sharp null hypothesis" that the treatment has no

effect on anyone (note that this is different from the usual null hypothesis of a zero average effect). If our original estimate is located far enough in the tails of the distribution created through re-randomization, we will reject the null.

Download and use the "ritest" package from Github: grantmcdermott/ritest: Randomization Inference on R model objects (github.com). Follow the examples and download instructions.

3. What causal effect is identified by the differences in means from questions 1 and 2 in the absence of cross-school externalities?

4. Suppose there are cross-school externalities from the deworming program. Discuss how externalities might bias our estimate from question 2.

5. Next, we will produce results similar to those in column (1) of table VII in Miguel and Kremer, the results of which are based on a regression of the following form:

$$any\_ics99_{ij} = \alpha + \beta_1 T_{1i} + X_{ij}^{'}\delta + \gamma_{03} N_{03i}^{T} + \gamma_{36} N_{36i}^{T} + \phi_{03} N_{03i} + \phi_{36} N_{36i} + e_{ij} \qquad (1)$$

Note that we will use a probit regression, just as Miguel and Kremer do, although for simplicity we have written equation (1) as a linear model.

In equation (1), any_ics99_{ij} is as defined above, T_{1i} is equal to 1 for students in Group 1 and 0 otherwise, N_{03i}^{T} is the number of pupils in schools assigned to Group 1 that are within a distance of three kilometers from school i, N_{36i}^{T} is defined similarly for distances between three and six kilometers, N_{03i} is the total number of pupils within three kilometers of school i, N_{36i} is defined similarly for distances between three and six kilometers, and e_{ij} is the error term. The pupil variables are given by `pop1_3km_original` etc. The covariate vector, $X_{ij}^{'}$, includes sap1 through sap4, Istd4 through Istd9, and mk96_s.

Estimate equation (1) using a probit model with clustered standard errors, weighting by *indiv_weight*. As we will learn later in the class, clustered standard errors allow for the regression error term to be correlated within (but not across) groups and allow for arbitrary heteroscedasticity. You will need to estimate clustered standard errors in a second step. I did this using the “vcovCL” function that comes with the “sandwich” package.

6. Explain why we will obtain biased estimates of γ_{03} and γ_{36} if we do not include N_{03i} and N_{36i} in our model.

7. Use the “margins” command (comes with Stata, available for installation in R) to estimate the following:

   - The average impact of being in a Group 1 school on *any_ics99*.
   - The average marginal effect on *any_ics99* of increasing the number of students within three kilometers who are in treated schools by 1,000 students.
   - The average marginal effect on *any_ics99* of increasing the number of students between three and six kilometers away who are in treated schools by 1,000 students.

   You should estimate all marginal effects holding other right hand side variables fixed at their sample averages. This is the default for “margins”.

   You have to put “margins” inside “summary” to see standard errors. I google searched for “margins vcovCL” to figure out how to do clustered standard errors with margins in R.

8. Now we will estimate our model allowing for a quadratic effect of the number of students in treated schools within a given distance. In other words, we will add $N_{di}^T \times N_{di}^T$ for each distance *d* to our model.

   Note that R uses different syntax for quadratic terms as opposed to interactions. To include the linear and quadratic versions of a variable x in your model, you would code `x + I(x^2)`. To include x, z, and their interaction in a model, you would code `x##z`.

Estimate a probit model of *any_ics99* that includes quadratic terms for the number of treated students with three kilometers and treated students between three and six kilometers away.

9. Compute the marginal effects for this new model and compare them to those obtained in question 6. You do not need to perform any hypothesis tests comparing the two sets of marginal effects. Just give your opinion as to whether our spillover estimates are sensitive to functional form.

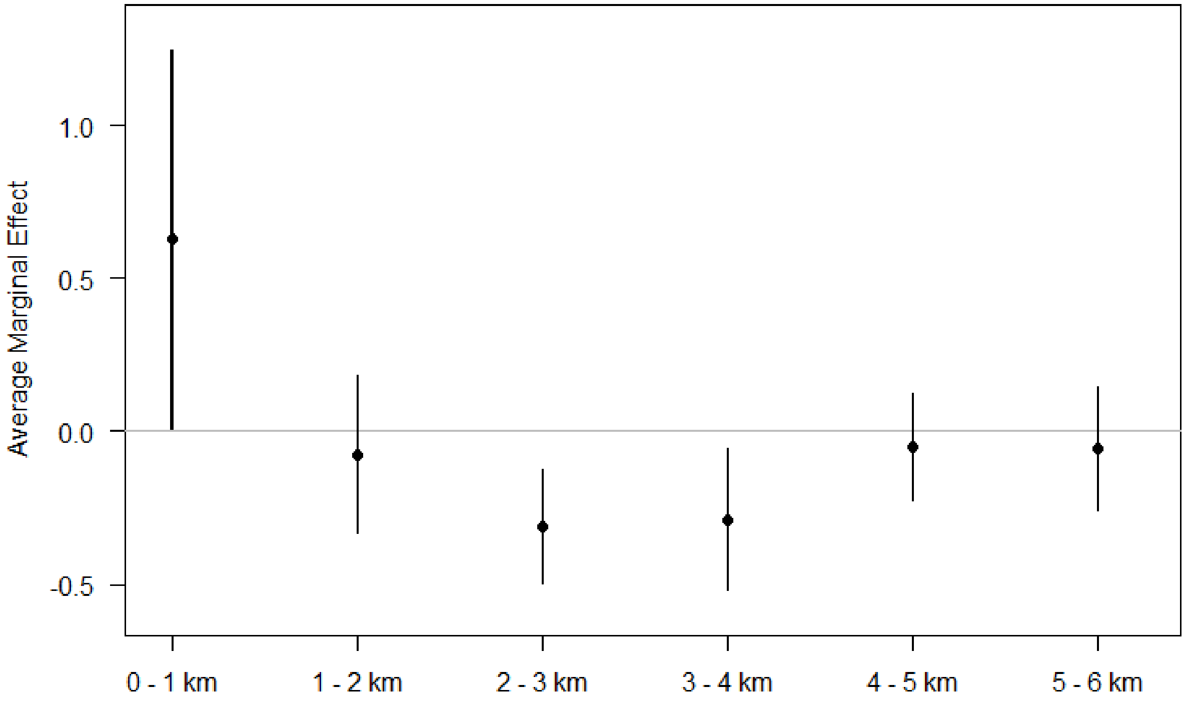
10. After a different set of authors did a replication of Miguel and Kremer that resulted in some doubt being cast on their results, both the original authors and the replicators conducted richer analyses of externalities using additional measures of the number of students in treated schools located within specified distances.

    The new variables use the naming conventions `pop1_0_1km_updated` (for number of treated students in treated schools between 0 and 1 kilometer away) and `popT_0_1km_updated` (for total number of students between 0 and 1 kilometer away). Estimate a new probit model that allows for externalities at distances of 0 to 1, 1 to 2, 2 to 3, 3 to 4, 4 to 5, and 5 to 6 kilometers away.

11. Estimate the marginal externality effect at each distance included in the new model. In R you must save the output from margins in order to answer the next question.

12. Use “plot.margins” in R to graph the complete set of marginal externality effects (see the “margins” documentation). Do you observe any patterns in the marginal effects? Does this pattern reflect what you would expect to observe?

    My graph in R looks as follows. I bet you can do better!


Average Marginal Effect
1.0
0.5
0.0
-0.5
0 - 1 km
1 - 2 km
2 - 3 km
3 - 4 km
4 - 5 km
5 - 6 km