REPLICATION EXERCISE 4: MIGUEL AND KREMER (2004)

VOJTĚCH BARTOŠ

In this exercise we try to get a better understanding and to replicate the paper by Miguel and Kremer (2004) on school deworming program in southern Busia, Kenya, and their impact on health and education. Read the introduction of the paper.

1. MK (2004): Paper Reading

(1) First, how do intestinal worms affect health status? And how do people get infected? Be specific about the two different types of worms the paper examines. (Section 2 might be of help for more thorough arguments)
(2) Second, most earlier studies have examined a link between education and health, but not vice versa. Why should we believe that health *causes* educational outcomes?
(3) Third, this study is not the first to examine the link between intestinal worms and the effects on individual health. Many medical journal articles were written on this topic.
   (a) Why should we, as economists, care about this link?
   (b) Why is the paper further (i.e. beyond the individual causal effect of worms on health and on education outcomes studied, e.g. by randomly providing kids within the same class with either a medicine or a placebo) interesting so that it got published in a top economics journal?
(4) What method does the paper use in order to estimate causal effects of worms on health and economic outcomes? It focuses on three different types of treatment effects. Which are these?
(5) Who exactly is eligible for treatment and who is not. Why?
(6) Describe the specific experimental design. On what level is the treatment distributed? Who serves as a control/comparison group? (See section 3)
(7) Describe the specifics of the treatment. What components does it have? This would be important in determining what effect are we actually capturing. (See section 3.2 specifically)
(8) How does the medical deworming treatment work specifically? How effective it is? Does it have a permanent effect or only temporary? Why is the paper focussing specifically on children? (See section 2)
(9) The pupils have been observed for an extended period of time and they might have moved from a treatment school to a control school and vice versa. When would that be a problem and what argument agains such problem do the authors make? (Section 3.3 and Table IV can help)

(10) Why can't the authors examine the within school externalities causally? How do they measure the effects of within-school externalities instead?
(11) How do the authors measure the between-school externalities? (in words)
(12) Explain how the models in equation (1) and (3) (pages 175 and 182, respectively) help us to estimate the desired effects of the deworming program on the 1) treated pupils, 2) non-treated pupils within the same school, and 3) non-treated pupils in surrounding schools. Explain briefly how to read the respective coefficients. This will be especially important for the empirical analysis when replicating Table VII.
(13) The paper focuses on two types of worm infections: schistosomiasis and geohelminth worms. What are the differences between the two in terms of transmission mechanisms? Why might the two infections have different predictions for the externalities within and across schools? (Section 2 can help)
(14) We won't get to that point, but what effect does the deworming treatment have on school attendance and on test scores? Comment and comment especially at the difference between the two results.
(15) The paper proposes to subsidize the deworming treatment. What is the argument for such a policy recommendation beyond the fact that the people might be liquidity constrained?
(16) The paper goes further and estimates a cost-benefit analysis of the deworming treatment. See how they estimate it if interested. This is beyond the scope of what we can cover today.

2. MK (2004): Data Work

Now we are in good shape to open the data. There are six datasets out of which we will use 5. These are:

- *comply.dta* — Data on pupils' deworming treatment status.
- *namelist.dta* — Data on school participation (attendance) of pupils, as recorded during visits by PSDP survey enumerators. Observations in this data set are for each visit for each pupil.
- *pupq.dta* — Data from 1998 and 1999 pupil questionnaires.
- *schoolvar.dta* — School-level data on zonal worm infection levels, 1996 district mock exam scores, pupil population and other characteristics for all 75 schools involved in the PSDP.
- *wormed.dta* — Data on helminth infections from 1998 and 1999 parasitological examinations, and hemoglobin concentrations in 1999.
- *test.dta*[1]

Identifiters, variable names:

- *pupid* — Throughout the data sets, pupils are identified by this seven digit identification number.

[1]The test.dta contains the data on test scores of the pupils. We won't get there due to the lack of time. But this dataset contains data from academic examination—ICS exams in 1998 and 1999, district mock exam in 1998, ICS drop out exam in 1998, and 1998 Kenya Certificate in Primary Education exam.

- *schid / schmk98 / sch98v1* — Primary schools are similarly tagged with three digit school identification codes which take various names in the data sets, but generally with the prefix "sch".
- *wgrp / wgrp1 / wgrp2 / wgrp3* — These are the group indicators; *wgrp* attains three values for Groups 1-3, the remaining variables are corresponding dummy variables for respective groups.

The estimation part is quite demanding. In the empirical analysis, we'll only go through the link between the deworming program and health outcomes. You can do the analysis on educational outcomes later, simply by replicating closely the remaining tables VIII-X.

(1) Replicating Table I
   (a) First thing we need to show is that the groups were similar prior to the intervention. This is what Table I does.
      (i) Why do we need to show this?
      (ii) We will use namelist and schoolvar datasets. Open both and familiarize yourself with the variables and the data in general.
   (b) Now have the namelist dataset open. Since we are interested in the pretreatment variables, we should restrict the sample to the earliest visit by dropping all observations from all later visits.
   (c) It seems that in the original paper there were some issues with duplicate observations. The authors detected these and marked them in variable *dupid*. Drop the duplicate observations.[2]
   (d) Now we merge the dataset with the pupq dataset. Use *pupid* as the unique identifier.
   (e) Create the following variables:
      (i) Share of days present in school in previous 4 weeks (they have 5 school days/week in Kenya) (see *absdays_98_6*).
      (ii) Child is often sick (see *fallsick_98_37*).
      (iii) Child is clean (see *clean_98_15*).
   (f) Read footnote a to Table I. The authors use school averages weighted by population. We want to replicate entire Panel A and the following variables of Panel B: attendance recored in school registers; Blood in stool; Child is often sick; Malaria; Child is clean. You can use `collapse` command in stata, the (mean) generates averages across groups (remember, by school), and (count) will generate the number of students in a particular school. When doing `summarize`, use analytical weights by number of pupils `aweight` (see help).
   (g) In order to examine the group difference, use a regression model that regresses the variable of interest on group treatments 1 and 2 (wgrp1 wgrp2). Again,

[2]This also means some differences in results, the authors actually admit this and we will encounter this on several occasions.

use analytical weights as in the step above (`aweight`), weight again by number of pupils per school.

(h) In order to replicate Panel C, we need to use the school level data in schoolvar.dta. We want to replicate the following variables: Distance from Lake Victoria; Pupil population; School latrines per pupil; Proportion moderate-heavy infections in zone; Group 1 pupils within 3km; Group 1 pupil within 3-6 km; Total primary school pupils within 3km; Total primary school pupils within 3-6 km. No need for weighting here, otherwise follow the same procedure as for Panels A and B.

(i) There is a specific reason for presenting Table I results in a paper using the method the authors use. What is it? Imagine a study with extremely large N (approaching infinity), would such a table be necessary to be presented from a theoretical point of view? Why (not)?

(j) Why do we care about the distance from Lake Victoria? It has to do with some specificities of the worm infections.

(k) Why do we care about the numbers of pupils within 3 and 3-6 km?

(2) Replicating Table III (only data on any medical treatment, not by the specific types of treatment).

(a) For this table we need both the *namelist* and the *comply* datasets. In order to be able to merge the data, let's first create and save the *comply* dataset in which we drop the duplicate IDs (use `duplicates` command in Stata, you want to drop duplicates in *pupid*).

(b) Now we load the *namelist* dataset, drop the duplicate IDs, we only keep the results for the first visit, and then we merge the data with the newly saved *comply* dataset (again, using *pupid*).

(c) Let's now do the summary statistics for receiving any deworming treatment in 1998 (*any98*) for students in grades (or "standard", *std*) 1 to 8 in early 1998, i.e. the first visit, by treatment and the 1998 eligibility status (*elg98*).

(d) Do the same analysis for any medical treatment in 1999 for grades 1 to 7 in early 1998, now for the eligibility variable for 1999.

(e) Do the same analysis for any medical treatment in 1999 for grades 1 to 7 in early 1998 among pupils "enrolled" in 1999.[3]

(f) In an ideal setting, how would the table look like? Write down the numbers in all cells, both for girls under 13 and boys, and for girls over 13.

(g) Why did we drop the grade 8 cohort from the sample in parts (d) and (e)?

(h) Was the treatment administered to all eligible individuals or not? In case the compliance with treatment was imperfect, why would that be a problem and how could one tackle it in terms of estimation strategy used. Could an imperfect compliance actually help their case? (Section 3.3 can help)

[3]Since the administrative data are noisy, the authors rely on the random checks in schools they conducted several times in each year of study. These are recorded in variables *totprs98* and *totprs99*. They consider a pupil ineligible if he or she was never present during the visit. (Note: also drop those for whom there are missing data on the school checks)

(i) There were also some cases of treatment in the control group. How was this possible? (Section 3.3 can help)
(j) Why is the rate of treatment lower among those eligible in the second year of the study? (Section 3.2 can help)

(3) Replicating Table V (only examining any moderate-heavy infections, anemia, and worm prevention behavior)
(a) Where do the data for table V come from (specifically, the *wormed* dataset)? Why don't we have data for Group 3?
(b) Load the *namelist* dataset, drop the duplicate IDs and keep only the first visit observations. Merge with the *wormed* dataset.
(c) Just keep the data for which we have all observations (i.e. `drop if _merge!=3`).
(d) Further, restrict the sample so that we have the data for moderate-heavy infections for 1999 for every individual, drop those with missing values.
(e) To reconstruct panel A, summarize moderate-heavy infections in 1998 and 1999 by treatment. Also, to reconstruct the differences between Groups 1 and 2, run a regression with moderate-heavy infections on LHS and group 1 indicator on the right hand side. Use Huber-White robust standard errors and cluster at school level (see footnote a for Table V).
(f) For panel B, we need to create an "anemia" dummy. The authors classify someone as anemic if hemoglobin level is below 100 grams per liter.[4]
(g) Now do the same anlaysis as above, now for the anemia dummy.
(h) In order to reconstruct Panel C, we need to load the *namelist* dataset again, drop the duplicate IDs and keep only the first visit observations. Merge with the *pupq* dataset.
(i) We need to create the variables. Create again the "child clean in 1999" indicator and additionally create an indicator for the child wearing any type of shoes in 1999 (i.e., shoes or slippers; see variable *shoes_99*). The "days contact with fresh water in past week" in 1999 is *dayswat_99_36*.
(j) Once you have the variables created, run the same analysis as in the case of Panels A and B.[5]
(k) Why didn't we simply merge the *pupid* dataset to the restricted sample we used for panels A and B?
(l) Briefly comment on the results in all three panels. What have we learned? Remember your response to (7) in the joint classroom discussion? What part of the two components of the treatment was most likely more effective?
(m) What do these results capture? Are we capturing the full effect of the treatment or are these results understating the true effect? Why?

(4) Replicating Table VI (just the moderate-heavy infection data for 1999)

[4]A quick check at Wikipedia suggests that "diagnosis in men is based on a hemoglobin of less than 130 to 140 g/L, while in women, it must be less than 120 to 130 g/L." So maybe the authors are a little too conservative here.

[5]For some reason, the authors now also use the Group 3. I would just drop them for consistency.

(a) Now we examine the role of externalities within schools. We discussed that we do not have an experimental manipulation, but the results might still be informative. Let's load the *namelist* dataset, drop the duplicate IDs and keep only the first visit observations. Merge it with the *wormed* and *comply* datasets.
(b) Restrict the sample to those with non-missing 1998 eligibility data (*elg98*) and to those with non-missing moderate-heavy infection data for 1999.
(c) Now we split the sample to eligible and non-eligible (remember, in Table V we pooled them). We'll only focus on the moderate-heavy infection results for 1999. Let's do the summary statistic for this variable for Group 1 treated in 1998, Group 1 untreated in 1998, Group 2 treated in 1999, and Group 2 untreated in 1999. Further, using the regressions we used e.g. in Table V, examine the differences in 1) Group 1 treated in 1998 and Group 2 treated in 1999, and 2) Group 1 untreated in 1998 - Group 2 untreated in 1999). Clustering at school level and using Huber-White robust standard errors. You should get the results as in Panel B, first row for girls under 13 and boys, and for girls over 13. Be careful about restricting the sample correctly for the regressions.
(d) Comment on the results briefly. What should results we expect had no within-school externalities be present?

(5) Replicating Table VII (remind yourself of the models we discussed in (12) of the paper reading part)
(a) Always store the results of all regressions using `outreg2` command (I am usually using the `dec(2) se` options to make the regression output more tractable).
(b) Load the *namelist* dataset, drop the duplicate IDs and keep only the first visit observations. Merge it with the *wormed*, *comply*, and *schoolvar* datasets.
(c) In order to be able to estimate model (3) from the paper, we need to create the variable for whether the individual actually received the treatment (*any98* / *any99*) when offered (in 1998 for Group 1 and in 1999 for Group2).
(d) We also need to create an interaction term variable of the variable we created in the previous step and the Group 1 indicator.
(e) If you want to get the same results as the authors, you can also divide the variables on Group 1 and total number of pupils within 3km and between 3-6km by 1000. (See the Table? Per 1000 pupils.)
(f) Now we are in a good shape to start building the regression models. We'll start with model (1). The authors estimate a probit model with moderate-heavy infection in 1999 on the left hand side and the following variables on the right hand side: *wgrp1 pop1_3km_original pop1_36k_original popT_3km_original popT_36k_original*. Do not include any controls at this point. Use marginal effects version, `dprobit`. Syntax is the same as for the regression, use Huber-White robust standard errors and cluster at the school level. Save the results and comment.

(g) Now we want to extend the model to replicate the model (3) in the paper. We'll just add the variables we created in points (c) and (d) above. Run the regression, save the results and comment.

(h) You can see that we only have about 2326 observations (two missing due to duplicates in the original data that we dropped). The reason is that the parasitological survey was conducted on a smaller sample. Hence, we should better try to weight the samples by the original school population. Let's do it.
   (i) First, save the current dataset, we'll need it right away.
   (ii) Now we need to get number of students per school. Remember, we did it already for Table I, point (1f) above. Just use the *numlist* dataset (dropped duplicates, using visit 1 data only) and do the `collapse / (count)`, by *schid*. This should give you 75 observations with school ID and number of pupils per school.
   (iii) Merge the data in memory with the data saved in part (i) above. Merge on school ID, we'll need to do a 1:m merge.
   (iv) In order to create proper weights, we need to adjust for the actual number of observations per school that we have in the original dataset saved in (i). Thus we need to know for how many students we actually have data used in the regression. You can do this using the following command: `egen ndata=count(pupid), by(schid)`.
   (v) To get the weights, let's just divide the actual number of pupils per school by the *ndata* variable created above.

(i) Now we are set to re-run the regressions we estimated in parts (f) and (g). Extend the syntax by including sampling weights based on the weighting variable we just created (`pweight`). Save the results and comment.

(j) The results still differ from MK substantially. We still did not include the controls.[6] These are the variables they add: *obs sap1 sap2 sap3 sap4 i.std mk96_s*. Enrich both models by this set of controls, run the regressions, save the results and comment.

(k) Now we have (almost) the same results as in Table V, columns (1) and (2). Let's now further replicate columns (4), (5), (7), and (8). We'll be examining the effects on schistosomiasis (*sm99_who*) and geohelminth (*any_geo99_original*) infections. Run the regressions, save the results and comment.

(l) Why should we expect the differential results on within/across school externalities for schistosomiasis and for the geohelminth prevalence? Are the results consistent with these predictions?

[6] See their description on p. 177 in the last paragraph of section 4.1