Problem Set 2

Impact Evaluation for Social Programs

December 2021

Due: December 22nd 2021, send to **charlotte.robert@awi.uni-heidelberg.de**

Required reading: *"Worms: Identifying impacts on education and health in the presence of treatment externalities", Miguel and Kramer, 2004*

The problem set will be graded out of 20 and the final grade is then standardized to be out of 10.

You can answer the Overview questions (Part 1) on your dofile or on a separate file. All tables created by your dofile should be included in your submission. You do-file should be attached to your submission.

Part 1 – Overview questions /10pts

Q1: (2pt) Read carefully "Worms: Identifying impacts on education and health in the presence of treatment externalities", Miguel and Kramer, 2004. What is the motivation and purpose of this paper? Describe the research question(s) that the authors attempt to answer, and explain why these questions are relevant/important.

relevant bc cheaper than other measures for school att. ; also relevant bc other studies did not find worm EXTERNAL effect. questions: are there ext effects of worm treatment on school attendance? and on test scores?

Q2: (1pt) Most earlier studies have examined a link between education and health, but not the other way around. Why should we believe that health causes educational outcomes?

bc there is lit that shows that and also in this paper they propose this casality

Q3: (1pt) Describe the specific experimental design. On what level is the treatment distributed? Who serves as a control/comparison group? (See section 3)

Q4: (3pt) The authors write that: "Given that there was no randomization of treatment within schools, Group 1 pupils who did not receive treatment in 1998 are compared to Group 2 pupils who did not receive treatment in 1999, the year that Group 2 schools were incorporated into treatment, to at least partially deal with potential bias due to selection into medical treatment." (p. 178) Explain, in your own words, why the authors are conducting such an analysis. How would selection into medical treatment threaten their results? What result do they find?

Q5: (3pt) The authors write that: "The larger [school] participation differences between treatment and comparison schools in 1998 may also have been due to the widespread El Nino flooding in this region in early 1998, which substantially increased worm loads between early 1998 and early 1999." (p.190) Explain how the increase in worm load in early 1998 due to El Nino could bias the results when comparing treatment and comparison schools. What would the direction of the bias be? Justify your answer.

Part 2 – Replicating results /10pts

Any attempt to replicate the analysis will be taken into consideration in the grading. Your do-file should be attached to your problem set submission and should be able to run on another computer than your own. **It is crucial that you report the number of the question you are answering in your do-file.**

You can choose to do I (Table I) or II (Table VI). If you choose to do both I and II, only I will be graded.

The datasets contained in the replication file 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

<u>Identifiters, variable names:</u>

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

- *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.

Since the replication of the empirical analysis is quite demanding, we will only replicate Table I or Table VI. The focus is thus only on the link between the deworming program and health outcomes. However, each table in Miguel and Kremer (2004) can be replicated using a single corresponding do file.

I - Replicating Table I – 10pt

(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) Open the *namelist* dataset. 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 using the variable *dupid*. Drop the duplicate observations. (This means that you will find some differences in results compared to the published tables; the authors admitted to this)

(d) Now merge the dataset with the *pupq* dataset. Use *pupid* as the unique identifier.

(e) Create the following variables:

(i) Share of days absent from 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 recorded in school registers; Blood in stool; Child is often sick; Malaria; Child is clean. You can use `collapse` command in stata, using the (mean) option to generate 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, 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) Present your result in a table (Excel or LaTeX).

(j) Comment on the results briefly.

II - Replicating Table VI (just the moderate-heavy infection data for 1999) – 10pt

(a) Let's examine the role of externalities within schools. Open the *namelist* dataset. 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.

(b) It seems that in the original paper there were some issues with duplicate observations. The authors detected these and marked them using the variable *dupid*. Drop the duplicate observations. (This means that you will find some differences in results compared to the published tables; the authors admitted to this)

(c) Now merge the dataset with the *wormed* and *comply* datasets. Use *pupid* as the unique identifier.

(d) Restrict the sample to those with non-missing 1998 eligibility data (*elg98*) and to those with non-missing moderate-heavy infection data for 1999.

(e) Now split the sample between eligible and non-eligible. We'll only focus on the moderate-heavy infection results for 1999 (Panel B, first line). 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. Following Table VI, 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). You should cluster errors at school level and use Huber-White robust standard errors (using `hreg`).

You should get the results as in Panel B, first row for girls under 13 and boys, and for girls over 13. Be careful that you restricted the sample correctly for the regressions.

(f) Present your result in a table (Excel or LaTeX).

(g) Comment on the results briefly.