FLS 6415: Class 6 Homework

September 28, 2017

Remember to answer all the questions in R markdown and produce a PDF. Email your completed homework (R markdown file and PDF) to jonnyphillips@gmail.com by midnight the night before class. Remember to refer to the example code from this week and the last couple of weeks for coding guidance.

This week we analyze the data from Reinikka and Svensson (2005), who use an instrumental variables methodology. The data comes from a Public Expenditure Tracking Survey in 2002, with each row of the data representing one school. Unfortunately, not all their data is publicly available, so we will not be able to exactly reproduce their findings. But we can use the available data from Uganda's 2002 Public Expenditure Tracking Survey to illustrate the use of instrumental variables.

NOTE: Reinikka and Svensson conduct a range of instrumental variable estimates in multiple papers, using different treatments and outcomes, but always the same instrument: distance to the nearest newspaper seller. Don't get confused - in this homework we're doing a slightly different analysis to the main Reinnika and Svensson paper but we're using the same instrument.

Table 1: Variables for Homework

Variable	Description
q11stud	Total Number of Students
q12p13	Number of Students in P1-P3
q13p47	Number of Students in P4-P7
q84agra	Total Cash Payment Actually Received
q29dist	Distance from the school to the Nearest Newspaper Seller
q30dist	Distance to the Nearest Bank Branch
q24teac	Number of Teachers
q25qual	Number of Qualified Teachers
q22ple	Number of students who took primary school exams
q23pass	Number of students who passed primary school exams

1. The **treatment** variable of interest is 'governance' or 'reduced corruption', specifically the proportion of grant funds intended for the school that is actually received. We need to create this variable. The numerator is simple: *q48agra* is the cash the school actually received. The denominator - what the school *should* have received - is based on a formula: 5,000 for every student in grades P1-P3 and 8,100 for every student in grades P4-P7. Calculate the treatment variable and provide a density plot that shows the distribution of this variable.

2. As your chart shows, there are infeasible outliers in the data. Remove any rows for which the treatment variable (% of grant received) is greater than 100%. Now produce a scatter plot of the amount of grant received (x-axis) against what the school should have got (y-axis). Add a 45-degree diagonal line through the origin to indicate what a perfect allocation would look like (where intended grant equals actual received grant). Hint: Use `geom_abline()` for the 45-degree line.

3. If we did not know about instrumental variables, the basic *observational* regression we might run is to examine directly how the treatment variable affects the outcome. First, calculate the outcome variable we are going to focus on: it is the **student-teacher ratio** (Total number of students/Number of teachers). Then run the regression of the outcome on treatment, and include a control for the total number of students in the school, just to control for a lot of variation due to school size. Interpret this regression.

4. Do you trust the causal effect estimates from Q3? What are the major threats to causal inference here? Provide concrete examples.

5. To conduct an Instrumental Variables Analysis, we first need to make sure we have a strong 'first stage', i.e. that our instrument (distance from the school to the nearest newspaper seller) predicts our treatment variable (% grant received). To copy Reinnika and Svensson we will use as the instrument `log(1+distance)` (we add one just to avoid taking the log of zero, which is undefined). So first create this instrument variable. Then conduct this first-stage regression, including the control for total number of students in the school again. Interpret the results (remembering how to interpret logs). Does this seem like a strong instrument?

6. The manual *2-Stage Least Squares* approach to an instrumental variables regression is to save the fitted values of the first stage regression from Q5 as another column in your data.frame. Then conduct the second-stage regression of the outcome on those fitted values, again including the control variable for the total number of students in the school. Interpret the Instrumental Variables regression results.

7. The only disadvantage of the 2-Stage Least Squares approach is that it might not accurately estimate the standard errors/uncertainty. Conduct the alternative all-in-one approach to conducting an instrumental variables regression using `ivreg` in the `AER` library. As before, the instrument is distance to the nearest newspaper seller, `log(1+distance)`, the treatment is the percentage of the grant received, and the outcome is the student:teacher ratio. Interpret the results and compare to the result in Q6.

8. The Instrumental Variables regression estimates the causal effect of an increase in governance, i.e. the % of grant received. If we wanted to ask a *different* causal question about the causal effect of distance to the nearest newspaper seller (i.e. the causal effect of the instrument, not of treatment), we can conduct the *reduced form* regression of the outcome on the instrument. Implement and interpret this regression.

9. A crucial assumption for the instrumental variables regression is the **exclusion restriction**: that the instrument ONLY affects the outcome through the treatment, and not through any other mechanism. We have to support this assumption by theory and qualitative evidence as it cannot be empirically verified. Make the case that distance to the nearest newspaper seller only affects the Student:Teacher ratio through its affect on the funding received.

10. Now pretend you are a reviewer/critic and make the case that the exclusion restriction assumption is likely to not be true.

11. One way of potentially improving the validity of the exclusion restriction is to develop an instrument in the first stage which focuses on the more 'exogenous' variation in distance. For example, adding a control for distance to the nearest bank branch creates an instrument that captures the idiosyncratic distance in the location of newspaper sellers *after* removing the distance to the nearest urban/economic centre. Add this control to both the first and second-stage and implement both the 2-Stage Least Squares methodology and the all-in-one methodology. Interpret these results and compare to your earlier analysis.

12. Using a `for` loop, `list`, and `stargazer`, construct a single output table that compares the Instrumental Variables estimates for three outcomes: (i) Student-Teacher Ratio (same as above), (ii) Percentage of qualified teachers (you need to calculate this; see the variables table above) and (iii) Percentage of students who passed primary school exams (you need to calculate this). Use the same treatment (% grant received), instrument (distance to nearest newspaper seller) as before, and include the control variables for total number of students (but not distance to the nearest bank branch). Interpret the direction, magnitude and significance for each outcome.