FLS 6415: Replication 7 - Controlling for Confounding

April 2020

To be submitted (code + answers) by midnight, Wednesday 6th May.

First read the paper by Baldwin (2013) on the course website.

The replication data is in the file *Baldwin_adjusted.csv*. A list of available variables is also provided below.

Variable	Description
connectionsMPchief	Number of years the MP has known the chief (already in 'log' form) - **Treatment**
tempclass08	Number of temporary classrooms in 2007-2008 - **Outcome**
tempclass07	Number of temporary classrooms in 2006-2007
pop2000	Village population in 2000
pop2009	Village population in 2009
experienceMP	Years since MP first elected
experiencechief	Years since Chief installed
voteMP06	% Vote for MP
MMD06	MP form governing party
diffvoteconst06	Difference in vote share between top two candidates
univMP	MP went to University
cabinetMP	MP has ever been in the cabinet
localMP	MP is from the chiefdom
secondaryedchief	Chief completed secondary education
politicalexpchief	Chief has ever participated in politics
agechief	Age of the chief in years
constcode	Constituency code
classneedper100	Students per Classroom 2006-07
yearinstalledchief	Year became chief
percturnout06	Turnout 2006 election
numcandidates	Number of candidates in 2006 election

1. We will focus on assessing Baldwin's (2013) claim that "politicians with stronger relationships to chiefs actually do provide more local public goods". Create a plot of the treatment variable (`connectionsMPchief`) against the outcome variable, the number of temporary classrooms in 2008 (`tempclass08`). Add a linear line of best fit to assess the relationship.

2. Implement the basic linear regression of the outcome on treatment with no controls/covariates. Interpret what you can conclude from this regression. *Note:* The `connectionsMPchief` variable is already in 'log' form (see Class 1 for guidance on how to interpret logged explanatory variables).

3. Provide two concrete, specific reasons for why our estimate in Q2 might be biased. In each case, which direction would the bias be?

4. Describe the Treatment Assignment Mechanism (why some units got treated and not others) for our treatment variable, the length of the relationship between MP and Chief.

5. Draw (by hand) the causal diagram (DAG) for our study, including the treatment effect of interest, the treatment assignment mechanism, and the threats to causal inference you described above. (Don't make it too complicated, just include the key variables and relationships!)

6. Based on your causal diagram (DAG) in Q5, apply the three rules we discussed about back-door paths and describe one set of control variables which would be sufficient to provide an unbiased estimate of the causal effect of treatment (if the DAG were correct).

7. One potential omitted variable (confounder) is population - in larger villages the MP and Chief are less likely to know each other personally, and village size might also affect the resources/demand for temporary classrooms. There are two potential control variables in the dataset we could use, a measure of population in 2000 (`pop2000`) and a measure of population in 2009 ('pop2009'). Which should we use, and why?

8. Run the simple linear regression of the outcome on treatment, controlling for any variables you identified as appropriate in Q6 and Q7 above. How do your results compare to the results of the regresssion in Q2?

9. Baldwin (2013) runs her regression using an ordered multinomial (ordered logit) model, reflecting the fact that the outcome variable (number of temporary classrooms) is not really continuous and can only take on a fixed set of integer values. Repeat your regression from Q8 but with an ordered logit model and interpret the results. (Note that Baldwin also clusters standard errors at the constituency (`constcode`) level, but don't worry about replicating this, just focus on the coefficient).

10. Now replicate the results from column (1) of Baldwin's Table 1, i.e. only include the control variables that she includes in Table 1. Compare the estimated treatment effect with your own model in Q9. (Again, don't worry about the standard errors).

11. Replicate all three columns of Table 1 in Baldwin (2013). How stable is the estimate of the treatment effect to alternative specifications of the control variables?