FLS 6415: Replication 7 - Controlling for Confounding

April 2020

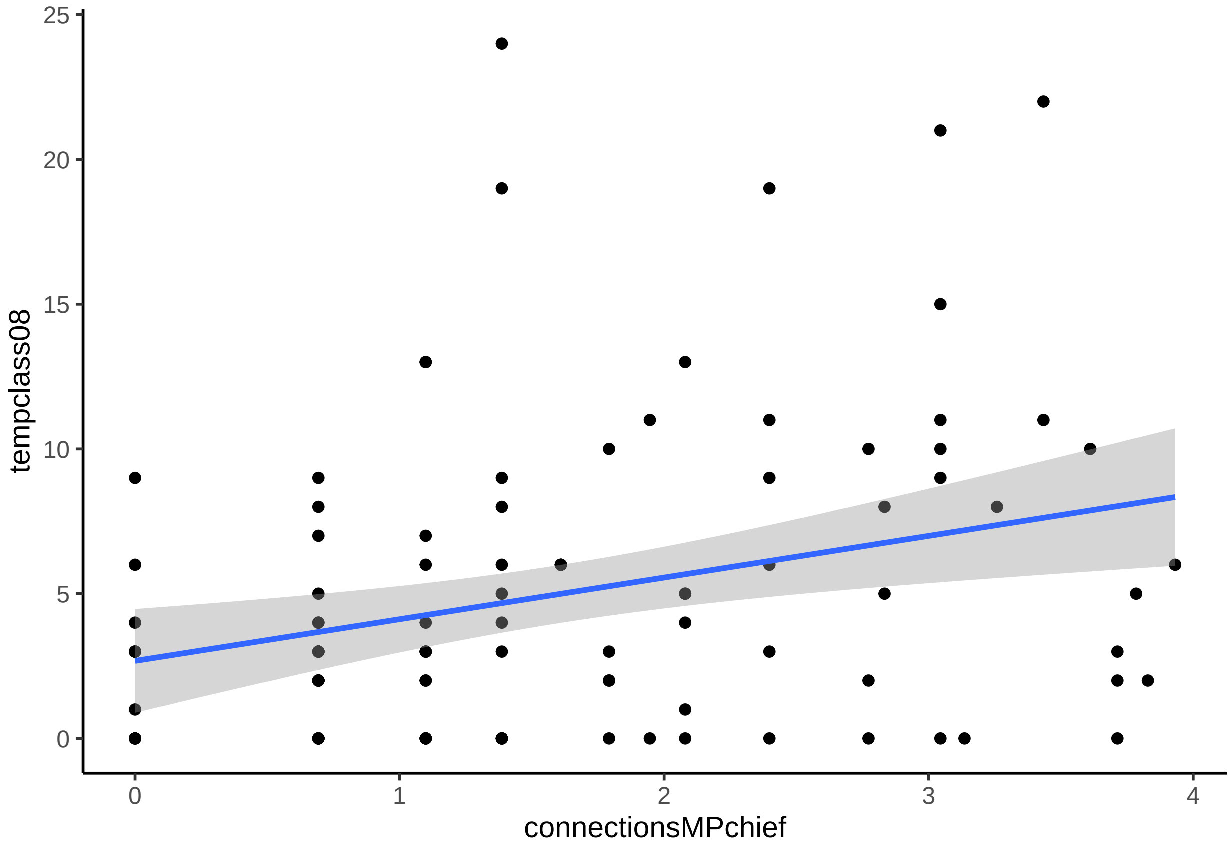
To be submitted (code + answers) by midnight, Wednesday 6th May.

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).**

Table 2: Q2

	Dependent variable:
	tempclass08
connectionsMPchief	1.439***
	(0.466)
Constant	2.679***
	(0.903)
Observations	101
R^2	0.088
Adjusted R^2	0.079
Residual Std. Error	5.073 (df = 99)
F Statistic	9.530*** (df = 1; 99)
Note:	*p<0.1; **p<0.05; ***p<0.01

A 1% increase in the number of years the chief has known the MP is associated with a 0.014 ($log(\frac{101}{100}) * 1.439$) increase in the number of temporary classrooms.

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

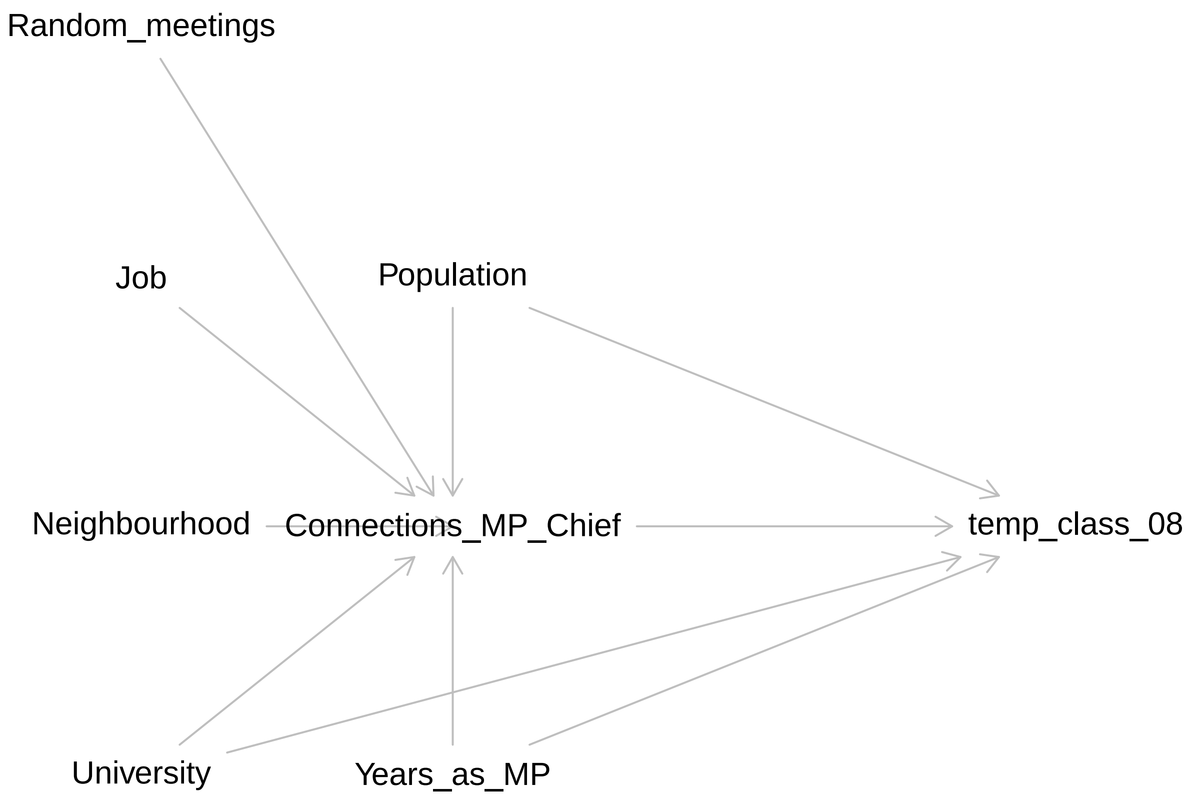
- The population of the village could be a confounding variable; population implies more children and could lead to the construction of more temporary classrooms (affecting our outcome variable), and population might make it harder for any single MP to know the chief personally (affecting the treatment varibale). So bigger villages would have low treatment and high outcome, while smaller villages would have high treatment and low outcome. This creates a 'false' negative correlation between treatment and outcome, which would lead us to *underestimate* the treatment effect.
- The number of years the MP has been elected may be a confounding variable. The longer the MP has been in office, the longer they are likely to have known an important figure such as the Chief (affecting the treatment); and we know from other research in political science that MPs that stay in longer for office usually face less competition and therefore might face weaker incentives to provide social services like temporary classrooms. Again, this creates a 'false' negative correlation between treatment and outcome, which would lead us to *underestimate* the treatment effect.

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.

Why do some MPs know the Chief for longer? That depends on how social networks are formed. Clearly there are some random/arbitrary elements like a chance encounter that became a friendship. But we know these networks are strongly influenced by socioeconomic conditions, by going to the same school or university, working in the same type of job, living in the same neighbourhood.

Note that some of these variables might plausibly affect the outcome (number of temporary classrooms) as well, which would make them confounders, for example whether the MP went to university. But some are unlikely to affect the outcome, for example the neighbourhhod they live in.

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).

The three back-door paths are:
1. Connections_MP_Chief <- Years_as_MP -> tempclass08
2. Connections_MP_Chief <- Population -> tempclass08
3. Connections_MP_Chief <- University -> tempclass08

This implies that the minimal set of control variables necessary to provide an accurate estimate of the causal effect is 'Years_as_MP', 'Population' and 'University'.

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?

We have to use `pop2000` to avoid post-treatment bias. It could be that better relations between chief and MP attract more people, increasing population, and therefore more classrooms are built. So controlling for population in 2009 might remove part of the real treatment effect.

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?

After controlling for the potential confounders, the estimate of the treatment effect declines.

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

Table 3: Q8

	Dependent variable:
	tempclass08
connectionsMPchief	1.143***
	(0.401)
univMP	0.559
	(0.937)
experienceMP	-0.055
	(0.104)
pop2000	0.142***
	(0.034)
Constant	0.280
	(1.058)
Observations	96
R^2	0.234
Adjusted R^2	0.200
Residual Std. Error	4.213 (df = 91)
F Statistic	6.933*** (df = 4; 91)
Note:	*p<0.1; **p<0.05; ***p<0.01

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).

Table 4: Q9

	Dependent variable:
	tempclass08
connectionsMPchief	0.443**
	(0.175)
univMP	0.088
	(0.391)
experienceMP	−0.043
	(0.045)
pop2000	0.062***
	(0.015)
Observations	96
Note:	*p<0.1; **p<0.05; ***p<0.01

A 1% increase in the number of years the chief has known the MP is associated with a 0.442% $(100 * (exp(log(101/100) * 0.443) - 1))$ increase in the odds of having one more temporary classroom.

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).

The coefficient of 0.352 is smaller than in the model in Q9. A 1% increase in the number of years the chief has known the MP is associated with a 0.351% $(100 * (exp(log(101/100) * 0.352) - 1))\%$ increase in the odds of having one more temporary classroom.

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?

The coefficients are relatively stable, providing confidence that the conclusion is not overly-sensitive to the specification.

Table 5: Q10

	Dependent variable:
	tempclass08
connectionsMPchief	0.352**
	(0.175)
tempclass07	0.385***
	(0.057)
pop2000	0.043***
	(0.015)
experienceMP	$-$0.057
	(0.042)
experiencechief	0.027**
	(0.013)
Observations	99
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Q11

	Dependent variable:		
	tempclass08		
	(1)	(2)	(3)
connectionsMPchief	0.352^{**}	0.378^{**}	0.331^{*}
	(0.175)	(0.181)	(0.184)
tempclass07	0.385^{***}	0.380^{***}	0.440^{***}
	(0.057)	(0.058)	(0.064)
pop2000	0.043^{***}	0.053^{***}	0.033^{*}
	(0.015)	(0.017)	(0.017)
experienceMP	-0.057	-0.054	-0.045
	(0.042)	(0.045)	(0.056)
experiencechief	0.027^{**}	0.023^{*}	0.009
	(0.013)	(0.013)	(0.018)
MMD06		-0.480	
		(0.563)	
voteMP06		0.932	
		(1.142)	
diffvoteconst06		-1.934	
		(1.371)	
univMP			-0.129
			(0.464)
cabinetMP			0.399
			(0.748)
localMP			0.482
			(0.478)
agechief			0.030^{*}
			(0.017)
secondaryedchief			-0.012
			(0.447)
politicalexpchief			0.617
			(0.508)
Observations	99	96	94
Note:	*p<0.1; **p<0.05; ***p<0.01		