FLS 6415: Class 2 Rough Homework Answers

August 31, 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.

First, some quick review questions about causal inference:

1. In one sentence, describe the fundamental problem of causal inference.

Causal inference requires us to measure the difference between potential outcomes **for the same unit**, but contrasting potential outcomes are defined in terms of treatment versus the absence of treatment, and only one of those conditions can be observed in reality.

2. We want to investigate the effect of attending university on support for redistribution. How might we define the counterfactual/control?

The control unit is very difficult to define because attending university takes many years and people could do many different things in that period that might also affect support for redistribution (eg. starting their own firm versus remaining unemployed). One option might be to take the 'median' activity for people who don't go to university, or simply leave it undefined and allow the units themselves to determine the control activity (in which the definition and interpretation of our treatment effect would need to be flexible).

3. Based on your knowledge of the Brazilian education system and society, describe the actual treatment assignment mechanism that applies to receiving the 'treatment' of going to university.

It's very complicated, involving a mix of assignments. Students themselves self-select into applying to university based on their interests, abilities, expectations and wealth. University availability also varies depending on geography and demand /supply. Exams then direct treatment to those who perform best in the exams, which involves some aspect of ability, some noise, but also a lot of socioeconomic background. That socioeconomic background is itself influenced by the accumulation of history, social organization and government policy. Policies such as quotas and redistributive support to particular groups may go someway to undoing these biases by altering the probability of treatment for, for example, afro-descendents.

4. The data below were collected from a randomized controlled trial that sent some high-school graduates to university but not others. By magic we have data on *both* potential outcomes for each student. Calculate (a) the average treatment effect, (b) the average treatment effect on the treated, and (c) the average treatment effect on the untreated:

Unit	D_i	Y_{1i}	Y_{0i}
A	0	7	6
B	1	8	4
C	0	6	4
D	1	6	6
E	0	4	5

The ATE is 1.2; the ATT is 2; and the ATU is 0.667.

5. Now imagine we only have observed outcomes and do not see both potential outcomes. Using only **observed** outcomes, calculate the estimated average treatment effect under the assumption that the data was generated from a randomized treatment assignment mechanism.

The observed average treatment effect is 2.

6. Is it surprising that the estimated average treatment effect in Q5 is different from your estimate of the average treatment effect in Q4? Why might they differ?

In Q5 we are estimating the real treatment effect using a subset of the data - since the unobserved potential outcomes are missing data - so due to random variation in who does or does not get treatment the result will vary. However, with a larger sample the observed treatment effect should converge to the real treatment effect.

7. We later find out that unit E ignored the randomized treatment assignment and was actually supposed to have taken the treatment, i.e. they should have gone to university but did not. Re-calculate the average treatment effect using the observed outcomes if Unit E had complied with the randomized treatment assignment and attended university.

If Unit E's treatment is changed, the estimated average treatment effect is 1.

8. As an estimate of the true causal effect of university on support for redistribution, do you trust your answer to Q5 or Q7 more? Why?

Q7's estimate is more accurate (it more closly represents the true average treatment effect) because treatment assignment in this case is random and therefore not subject to bias. Q5 provides an over-estimate because unit E who *loses* from treatment tries to avoid treatment, biasing the distribution of potential outcomes (this is self-selection). Therefore we observe unit E's'high' Y_0 rather than its 'low' Y_1 and we get an over-estimate of the average treatment effect.

Part 2

The second part of the homework is another exercise to increase your fluency in R. It focuses on the political economy causal question of whether institutions cause growth using data from Glaeser et al 2004, "Do Institutions Cause Growth?". The dataset is an incomplete panel covering 127 countries for four decades from 1960 - 2000. The measure of 'institutions' is "Initial executive constraints", a measure of checks and balances on government.

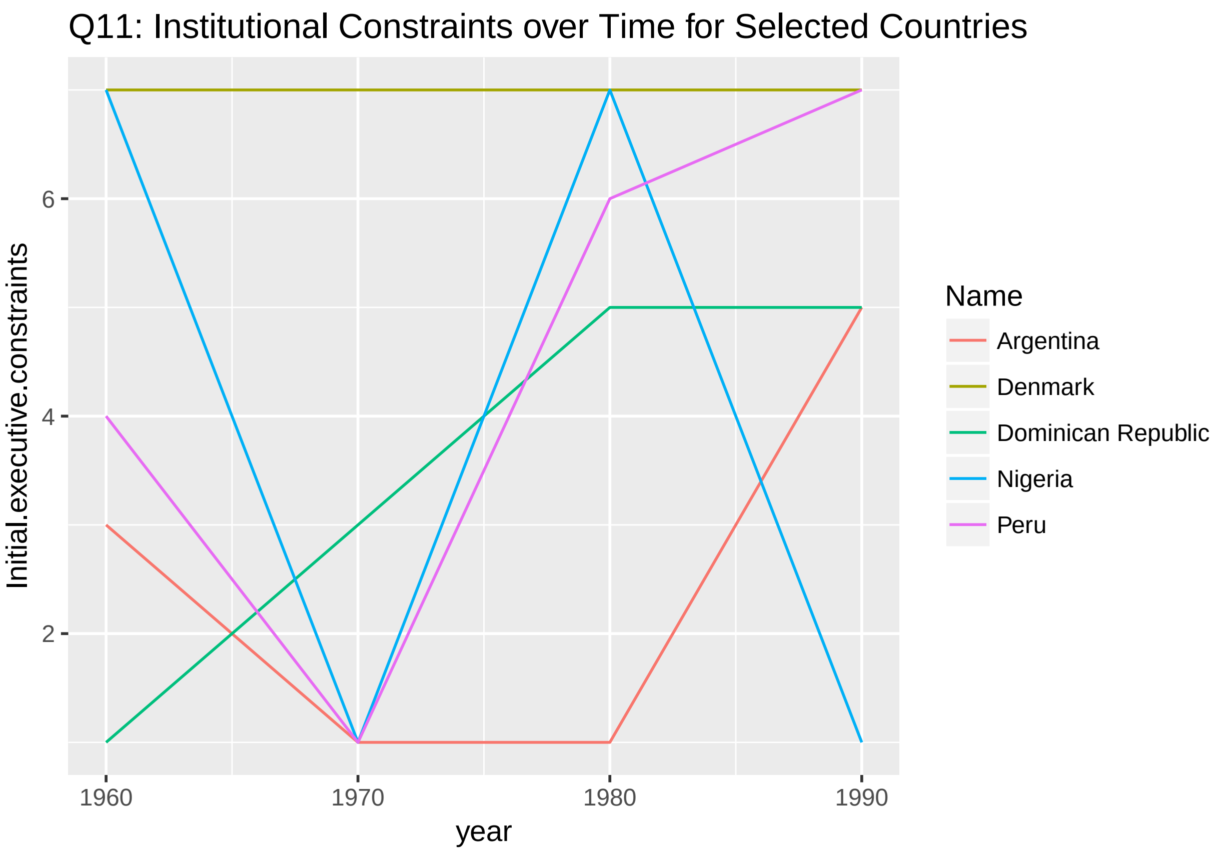
9. Download the XLS dataset "Do Institutions Cause Growth?" from http://faculty.tuck.dartmouth.edu/rafael-laporta/research-publications. Save the "data for table 5" sheet of the XLS file as a CSV. Start a new R markdown file and import the CSV.

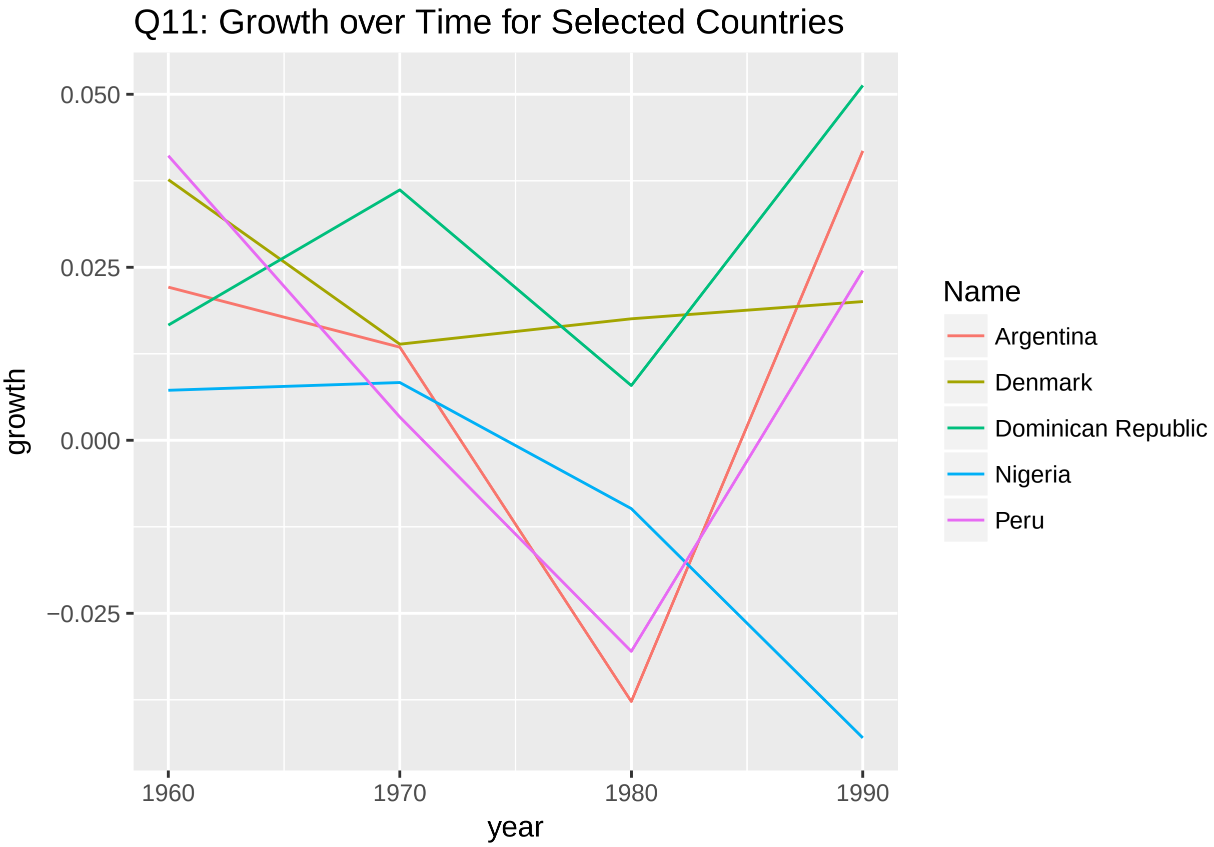
10. Create a simple table showing the average global growth rate for each decade (simple average across countries).

Table 2: Q10: Average growth (%) by decade

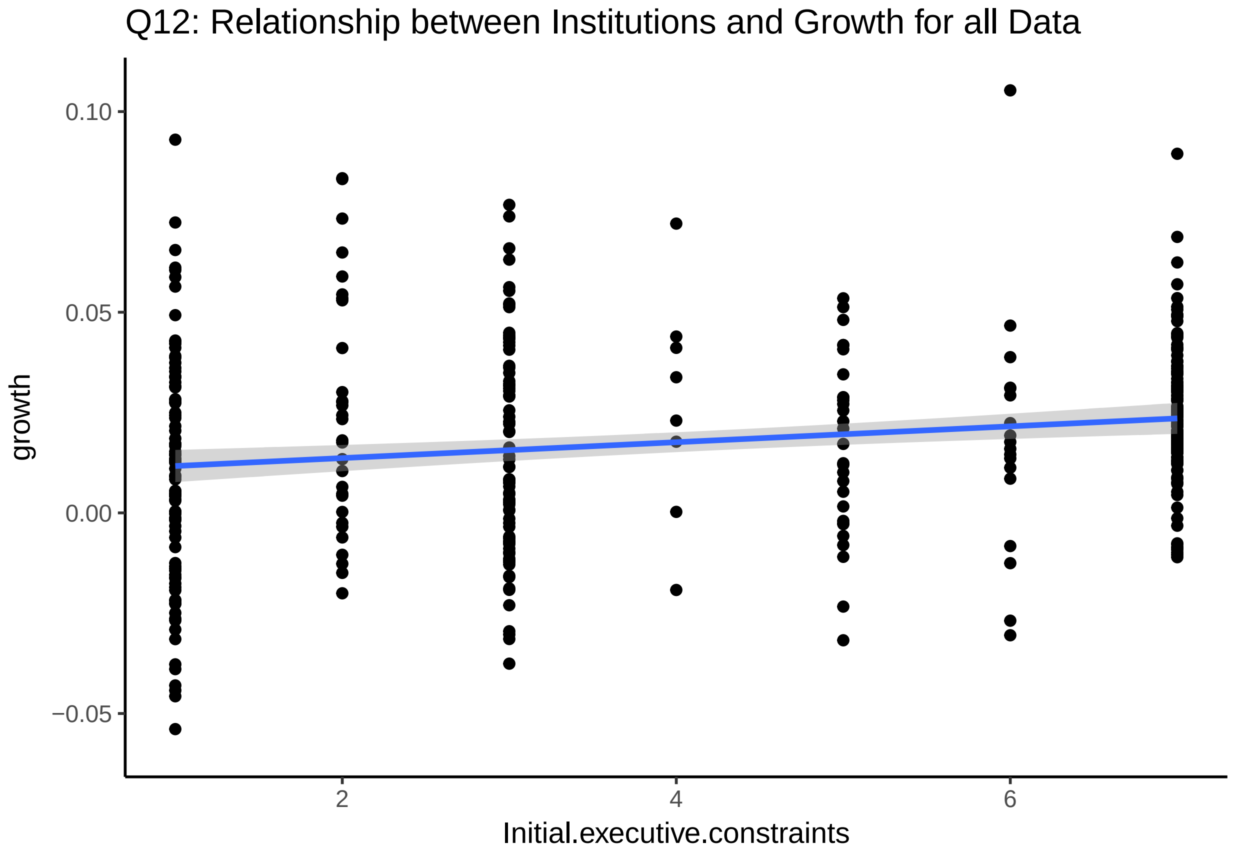
year	growth
1960	2.908
1970	2.101
1980	1.080
1990	1.503

11. Create a simple (labelled) chart showing institutions (executive constraints) over time for the following countries: Argentina, Denmark, Dominican Republic, Nigeria and Peru. Create a separate chart showing growth over time for the same countries.


Q11: Institutional Constraints over Time for Selected Countries
Initial.executive.constraints
6
4
2
1960
1970
1980
1990
year
Name
Argentina
Denmark
Dominican Republic
Nigeria
Peru



12. Create a simple (labelled) chart plotting institutions against growth for all the data points in the dataset. Add a linear line of best fit.



13. Conduct a linear regression of the following form: Growth ~ Institutions. Report the table of variables, coefficients and p-values. Interpret the coefficient and p-value on institutions.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0097	0.0025	3.93	0.0001
Initial.executive.constraints	0.0020	0.0005	3.81	0.0002

Table 3: Q13: Regression of Growth on Institutional Constraints

The average growth rate among countries which score zero on the executive constraints measure is 0.97%. Each additional unit increase in the executive constraints measures is associated with a 0.2% increase in growth. This effect is statistically significant at the p=0.05 level.

14. We might be worried about an omitted variable that explains both growth and institutions. Add one at a time the following variables to the regression: Log initial GDP per capita, Share of population living in temperate zone, Log initial years of schooling. How does each control variable change the interpretation of the regression results? What happens if you include all the control variables in the same regression?

Education (years of schooling) and temperate climate (share of population in the temperate zone) are positively associated with growth. In the combined regression initial GDP per capita is associated with a lower growth rate, which indicates (conditional) convergence so that initially poorer countries grow faster. While the direction of the executive constraints coefficient is consistently positive, its magnitude an statistical significance vary considerably depending on the controls we include - with the inclusion of education and temperate climate alone institutions are no longer associated with an increase in growth. In the combined regression, there is a smaller but statistically significant effect.

15. We might be worried that ‘treatment’ here is not independent because we’re repeatedly measuring the

Table 4: Q14: Regressions of Growth on Institutional Constraints with Controls

	Dependent variable:			
	growth			
	(1)	(2)	(3)	(4)
Initial.executive.constraints	0.002** (0.001)	0.0003 (0.001)	0.0005 (0.001)	0.001* (0.001)
Log.initial.GDP.per.capita	0.002 (0.002)			$-$0.013*** (0.002)
Log.initial.years.of.schooling		0.006*** (0.002)		0.009*** (0.002)
Share.of.population.living.in.temperate.zone..1995.			0.018*** (0.003)	0.024*** (0.004)
Constant	$-$0.003 (0.011)	0.010*** (0.003)	0.009*** (0.002)	0.097*** (0.017)
Observations	382	315	372	308
R^2	0.040	0.075	0.112	0.188
Adjusted R^2	0.035	0.069	0.107	0.177
Residual Std. Error	0.025 (df = 379)	0.023 (df = 312)	0.024 (df = 369)	0.021 (df = 303)
F Statistic	7.887*** (df = 2; 379)	12.580*** (df = 2; 312)	23.312*** (df = 2; 369)	17.552*** (df = 4; 303)
Note:				*p<0.1; **p<0.05; ***p<0.01

same country. Add fixed effects (dummy variables) for each country to the regression (excluding the control variables). How does this change the interpretation of the regression results?

% latex table generated in R 3.3.3 by xtable 1.8-2 package % Fri Sep 01 09:18:22 2017

With the inclusion of country fixed effects to take account of country-specific differences in growth levels, institutions no longer have a positive or statistically significant effect on growth.

16. (Slightly harder) We might be worried about reverse causation so that institutions explain growth. One rough way of testing this is to see whether growth in, for example, the decade 1960 is related to subsequent institutions in 1970. Adjust your dataset so that each row contains the growth for a particular decade with the institutions at the *end* of that decade, and conduct a regression of Institutions ~ Growth. (Hint: Use the 'mutate' function in dplyr to reorganize the data).

The regression of institutions on growth in the preceding decade indicates that there is a positive, significant and substantial effect, which suggests that reverse causation is a possibility.

17. Do the regressions in Q13 (and Q16) provide estimates of causal effects? Why or why not?

None of these regressions estimates causal effects because the treatment assignment mechanism is likely to be complex and systematically biased by omitted variables, self-selection or reverse causation.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0030	0.0215	0.14	0.8885
Initial.executive.constraints	-0.0007	0.0009	-0.80	0.4263
as.factor(worldbankcode)ARG	0.0088	0.0239	0.37	0.7139
as.factor(worldbankcode)AUS	0.0241	0.0242	1.00	0.3186
as.factor(worldbankcode)AUT	0.0314	0.0242	1.30	0.1948
as.factor(worldbankcode)BDI	-0.0187	0.0247	-0.76	0.4508
as.factor(worldbankcode)BEL	0.0302	0.0242	1.25	0.2122
as.factor(worldbankcode)BEN	-0.0037	0.0247	-0.15	0.8797
as.factor(worldbankcode)BFA	0.0039	0.0239	0.16	0.8709
as.factor(worldbankcode)BGD	0.0262	0.0262	1.00	0.3174
as.factor(worldbankcode)BOL	0.0032	0.0239	0.13	0.8946
as.factor(worldbankcode)BRA	0.0269	0.0239	1.13	0.2613
as.factor(worldbankcode)BWA	0.0769	0.0264	2.92	0.0038
as.factor(worldbankcode)CAF	-0.0173	0.0247	-0.70	0.4851
as.factor(worldbankcode)CAN	0.0259	0.0242	1.07	0.2840
as.factor(worldbankcode)CHE	0.0166	0.0242	0.69	0.4935
as.factor(worldbankcode)CHL	0.0240	0.0239	1.01	0.3157
as.factor(worldbankcode)CHN	0.0413	0.0239	1.73	0.0845
as.factor(worldbankcode)CIV	0.0017	0.0239	0.07	0.9434
as.factor(worldbankcode)CMR	0.0035	0.0239	0.15	0.8822
as.factor(worldbankcode)COG	0.0338	0.0239	1.42	0.1579
as.factor(worldbankcode)COL	0.0204	0.0240	0.85	0.3974
as.factor(worldbankcode)COM	-0.0136	0.0262	-0.52	0.6045
as.factor(worldbankcode)CRI	0.0153	0.0242	0.63	0.5285
as.factor(worldbankcode)CYP	0.0499	0.0249	2.00	0.0462
as.factor(worldbankcode)DEU	0.0229	0.0249	0.92	0.3588
as.factor(worldbankcode)DNK	0.0245	0.0242	1.01	0.3116
as.factor(worldbankcode)DOM	0.0276	0.0239	1.16	0.2487
as.factor(worldbankcode)DZA	0.0101	0.0247	0.41	0.6830
as.factor(worldbankcode)ECU	0.0144	0.0239	0.60	0.5469
as.factor(worldbankcode)EGY	0.0253	0.0239	1.06	0.2908
as.factor(worldbankcode)ESP	0.0337	0.0239	1.41	0.1601
as.factor(worldbankcode)ETH	0.0030	0.0239	0.13	0.9000
as.factor(worldbankcode)FIN	0.0313	0.0242	1.29	0.1965
as.factor(worldbankcode)FJI	0.0190	0.0264	0.72	0.4715
as.factor(worldbankcode)FRA	0.0267	0.0239	1.11	0.2659
as.factor(worldbankcode)GAB	0.0350	0.0247	1.42	0.1575
as.factor(worldbankcode)GBR	0.0229	0.0242	0.95	0.3433
as.factor(worldbankcode)GHA	0.0113	0.0239	0.47	0.6360
as.factor(worldbankcode)GIN	-0.0016	0.0239	-0.07	0.9481
as.factor(worldbankcode)GMB	0.0030	0.0247	0.12	0.9026
as.factor(worldbankcode)GNB	0.0200	0.0261	0.77	0.4440
as.factor(worldbankcode)GNQ	-0.0040	0.0247	-0.16	0.8721
as.factor(worldbankcode)GRC	0.0322	0.0239	1.34	0.1805
as.factor(worldbankcode)GTM	0.0119	0.0239	0.50	0.6177
as.factor(worldbankcode)GUY	-0.0070	0.0262	-0.27	0.7902
as.factor(worldbankcode)HND	0.0066	0.0247	0.27	0.7890
as.factor(worldbankcode)HTI	-0.0093	0.0247	-0.38	0.7073
as.factor(worldbankcode)HUN	0.0224	0.0247	0.91	0.3655
as.factor(worldbankcode)IDN	0.0320	0.0239	1.34	0.1820
as.factor(worldbankcode)IND	0.0294	0.0242	1.21	0.2256
as.factor(worldbankcode)IRL	0.0428	0.0242	1.77	0.0779
as.factor(worldbankcode)IRN	0.0426	0.0302	1.41	0.1595
as.factor(worldbankcode)ISL	0.0302	0.0242	1.25	0.2131
as.factor(worldbankcode)ISR	0.0310	0.0242	1.28	0.2010
as.factor(worldbankcode)ITA	0.0313	0.0242	1.29	0.1968
as.factor(worldbankcode)JAM	0.0014	0.0249	0.05	0.9562
as.factor(worldbankcode)JOR	0.0116	0.0239	0.48	0.6285
as.factor(worldbankcode)JPN	0.0439	0.0242	1.82	0.0704

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.8312	0.1801	21.27	0.0000
growth	14.4799	5.6647	2.56	0.0111

Table 6: Q16: Regression of Institutions on Growth in Previous Decade