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.

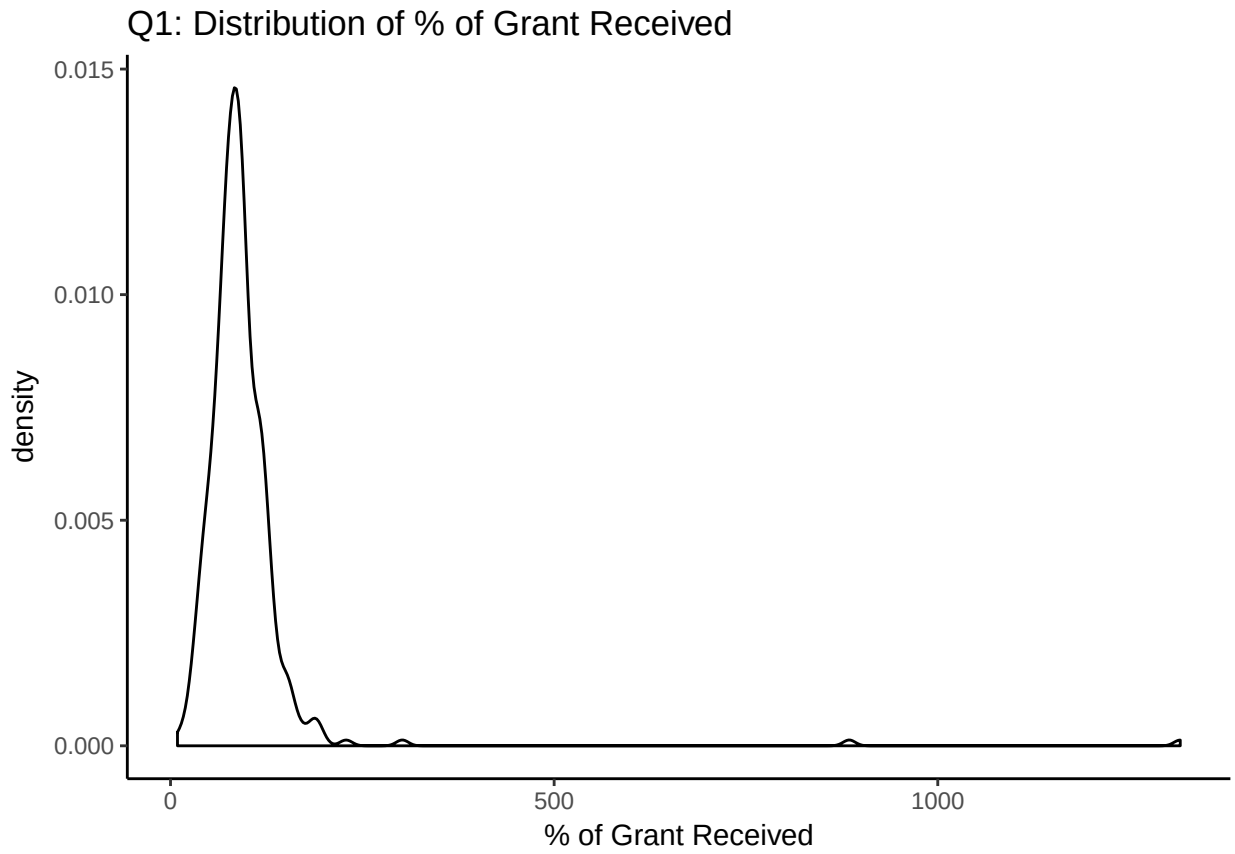
This week we analyze the data from Reinikka and Svensson (2005), who use an instrumental variables methodology. The data comes from a Public Expenditure Tracking Survey in 2002, with each row of the data representing one school. Unfortunately, not all their data is publicly available, so we will not 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

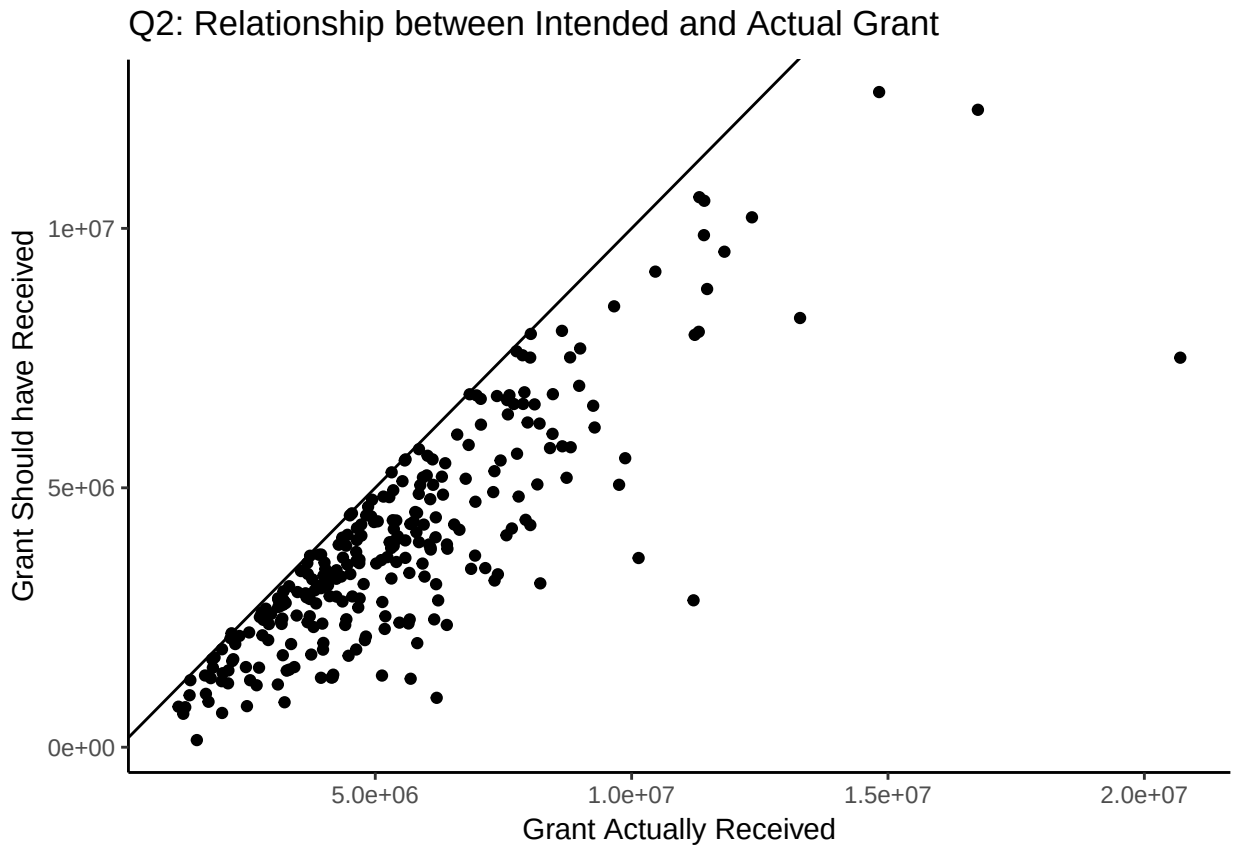
Variable	Description
q11stud	Total Number of Students
q12p13	Number of Students in P1-P3
q13p47	Number of Students in P4-P7
q84agra	Total Cash Payment Actually Received
q29dist	Distance from the school to the Nearest Newspaper Seller
q30dist	Distance to the Nearest Bank Branch
q24teac	Number of Teachers
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.

Note: The question should probably have swapped the x and y variables to make the chart look better, but it's the same thing either way round.



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.

	term	estimate	std.error	statistic	p.value
1	(Intercept)	58.43	3.83	15.25	0.00
2	pct_grant_received	-0.13	0.04	-2.87	0.00
3	q11stud	0.01	0.00	4.42	0.00

Table 2: Q3: Observational Regression of Student:Teacher Ratio on Percentage Grant Received

Controlling for the size of the school, a 1% point increase in the share of its intended grant a school receives is on average associated with a reduction in the student:teacher ratio of 0.13 (0.13 fewer students per teacher). This suggests that reduced corruption is associated with smaller (and therefore more effective) class sizes.

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

This is not a causal effect estimate since the regression is purely observational and potential outcomes are unlikely to be balanced. The major threats to causal inference here include reverse causation and confounding. For example, poor districts may experience higher rates of corruption and also have higher student-teacher ratios if there are fewer available qualified teachers. This would produce a significant coefficient in the observational regression even if there were no causal effect of corruption on student-teacher ratios.

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?

	term	estimate	std.error	statistic	p.value
1	(Intercept)	79.97	4.11	19.45	0.00
2	log(1 + q29dist)	-2.07	1.10	-1.88	0.06
3	q11stud	-0.00	0.00	-1.24	0.22

Table 3: Q5: First-Stage Regression of Percentage Grant Received on Distance to Nearest Newspaper Seller

Controlling for school size, a 1% increase in distance from the newspaper seller is on average associated with 0.02% points less of the intended grant being received. However, the p-value on this relationship is borderline significant and the F-statistic is well below 10, suggesting that this is a very weak 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.

```
d$Fitted_2SLS <- lm(as.formula(paste(treatment,"~",instrument,controls)),data=d)$fitted.values
regQ6 <- tidy(lm(as.formula(paste(outcome,"~ Fitted_2SLS",controls)),data=d))

print(xtable(regQ6,caption="Q6: 2SLS Instrumental Variables Regression"),comment=FALSE)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	146.24	28.90	5.06	0.00
2	Fitted_2SLS	-1.31	0.39	-3.38	0.00
3	q11stud	0.01	0.00	3.04	0.00

Table 4: Q6: 2SLS Instrumental Variables Regression

Controlling for school size, the instrumental variables estimate indicates that a 1% point increase in the share of the intended grant that a school receives is on average associated with 1.31 fewer students for every teacher in the school. That is ten times the magnitude of the effect estimated from the observational regerssion in Q3. This estimate is statistically significant at the 1% level.

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.

```
regQ7 <-summary(ivreg(as.formula(paste(outcome,"~",treatment,controls, "|",instrument,controls)),data=d
print(xtable(coefficients(regQ7),caption="Q7: IVREG Instrumental Variables Regression"),comment=FALSE)
```

The coefficient estimate and interpretation is identical to in Q6. The only difference is that the standard errors are nearly doubled and the p-value is now only significant at the 10% level.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	146.24	55.45	2.64	0.01
pct_grant_received	-1.31	0.75	-1.76	0.08
q11stud	0.01	0.00	1.58	0.11

Table 5: Q7: IVREG Instrumental Variables Regression

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.

```
regQ8 <- tidy(lm(as.formula(paste(outcome,"~",instrument,controls)),data=d))
print(xtable(regQ8,caption="Q8: Reduced Form Regression"),comment=FALSE)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	41.12	3.00	13.69	0.00
2	log(1 + q29dist)	2.72	0.81	3.38	0.00
3	q11stud	0.01	0.00	5.35	0.00

Table 6: Q8: Reduced Form Regression

A 1% increase in a school's distance from a newspaper seller is on average associated with a 0.027 increase in the student-teacher ratio. This is statistically significant.

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.

It is true that the distance between a school and a newspaper seller can affect the student-teacher ratio. But the only effect this this distance is likely to have is by making more resources available to the school, and this is exactly the instrumental variables pathway we are seeking to use. While it is impossible to prove that distance has no other effects, consider that the decisions of where to sell newspapers are idiosyncratic, based on hundreds of individuals' business decisions and without reference to the overcrowding situation in local schools. Moreover, schools are fixed structures that cannot be moved and so there cannot be any strategic manipulation of this distane by actors who might also influence the student-teacher ratio. In short, any effect of distance is only likely to impact student-teacher ratio throught the availability of resources that ultimately determine this ratio; there are few other plausible pathways.

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

There are many pathways by which distance between a school and the nearest newspaper seller could affect the student-teacher ratio, and not all of these work solely through a change in corruption or resources passed to the school. Consider, for example, that teachers like to read the newspaper. When posted to schools far from newspaper sellers, teachers rapidly abandon their job, which immediately leads to an increase in the student-teacher ratio. This direct relationship exists even if there is no change in the share of grant resources that actually reach the school. Many other factors could also produce correlation between the instrument and treatment, for example poorer and more rural areas that tend to be further from newspaper sellers may be populated by parents who are more likely to send their child to public school than in richer urban areas where children are sent to private school, lowering stuent-teacher ratios regardless of the proportion of resources that reach the schools.

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.

```
controls <- "+ q11stud + log(1+q30dist)"
regQ11a <- ivreg(as.formula(paste(outcome,"~",treatment,controls, "|",instrument,controls)),data=d)

#print(xtable(coefficients(regQ11a),caption="Q11: IVREG Regression with Control for Distance to Bank"),

d$Fitted_2SLS <- lm(as.formula(paste(treatment,"~",instrument,controls)),data=d)$fitted.values
regQ11b <- lm(as.formula(paste(outcome,"~ Fitted_2SLS",controls)),data=d) #Should be negative

#print(xtable(regQ11b,caption="Q11: Two-Stage Least Squares Regression with Control for Distance to Ban

stargazer(regQ11a,regQ11b,title="Q11: Two-Stage Least Squares Regression with Control for Distance to Ba
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Oct 05, 2017 - 7:56:07 AM

Table 7: Q11: Two-Stage Least Squares Regression with Control for Distance to Bank

	Dependent variable:	
	IVREG	2SLS
	instrumental variable	*OLS*
	(1)	(2)
pct_grant_received	−0.475	
	(0.638)	
Fitted_2SLS		−0.475
		(0.571)
q11stud	0.011***	0.011***
	(0.003)	(0.003)
log(1 + q30dist)	3.682**	3.682***
	(1.538)	(1.375)
Constant	71.531	71.531
	(51.525)	(46.059)
Observations	269	269
R^2	−0.046	0.164
Adjusted R^2	−0.058	0.155
Residual Std. Error (df = 265)	14.954	13.368
F Statistic		17.336*** (df = 3; 265)
Note:		*p<0.1; **p<0.05; ***p<0.01

Both methodologies suggest that after controlling for distance to the bank and using the instrument of distance to the nearest newspaper seller, the effect of receiving a 1% greater share of the intended grant is associated with a 0.475 reduction in the student-teacher ratio, but this difference is not statistically significant.

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.

Table 8: Instrumental Variables Regressions of Outcomes on Treatment with Controls

	STR	**pct_qual_teachers**	**pct_exam_pass**
pct_grant_received	-1.314^{*}	1.878	2.755
	(0.747)	(1.168)	(2.066)
q11stud	0.007	0.014^{**}	0.032^{***}
	(0.004)	(0.007)	(0.009)
Constant	146.238^{***}	−58.299	−165.551
	(55.446)	(86.624)	(152.604)
N	269	269	250
R-squared	−2.283	−4.829	−3.228
Adj. R-squared	−2.307	−4.873	−3.262
Residual Std. Error	26.440 (df = 266)	41.308 (df = 266)	55.956 (df = 247)

***p < .01; **p < .05; *p < .1

Controlling for school size, a 1% increase in the share of the grant received is on average associated with 1.3 fewer students for every teacher in the school, a 1.8% point increase in the proportion of teachers who are qualified, and a 2.8% point increase in the student exam pass rate. However, only the change in the student-teacher ratio is statistically significant at the 5% level.