Instrumental Variables Lab

Getting started

1. Load the `Smoking.dta` data file. This is a very small dataset containing information about cigarette consumption, prices, and taxes for 48 U.S. states.

2. Start by summarizing the variables using tables or graphs.

3. Create logged transformations of some key variables:
   ```
   gen logpacks = log(packs)
   gen logrprice = log(rprice)
   gen logincome = log(income)
   ```
   You should use the logged forms of these variables in the following analyses.

4. Our goal is to estimate the effect of cigarette prices (`logrprice`) on smoking (`logpacks`). Construct an OLS estimate of this effect.

Finding instruments

5. State-level cigarette prices and smoking rates may be influenced by many things. Draw a causal graph of this causal relationship along with any confounding variables. Where would you put an instrumental variable to help identify an effect?

6. Some authors have suggested state-level cigarette tax rates (`tax`) could be a valid instrument. Does this seem like a *credibly exogenous* instrument for this problem?

7. Others have suggested that a better instrument is the amount of VAT (general sales tax on all purchased goods; `tdiff`) is a better instrument. Does this seem like a *credibly exogenous* instrument? Is it more or less so than cigarette-specific taxes?

8. Test whether cigarette-specific and general tax rates are *relevant* instruments.

Two-stage least squares estimation

9. Let's assume that `tdiff` is a credible instrument. Manually generate a two-stage least squares estimate by regressing logged cigarette prices on `tdiff`. Based on this first stage equation, do general sales taxes seem to be a strong instrument?

10. Save the fitted values from this regression as a new variable `pricefit`.

11. Manually estimate the second stage equation using `pricefit` to predict `logpacks`. How does the estimate compare to your OLS estimate from above?

12. Remember that the SEs calculated in this method are incorrect. To obtain correct SEs, we need to use the `ivregress` command. It has the following syntax:
    ```
    ivregress 2sls outcome [exogenousvarlist] (endogenousvarlist = ivlist)
    ```
    How do coefficient estimates and SEs from `ivregress` compare to the estimates from manually performing two-stage least squares?

First-stage diagnostics

13. To include additional details in the output, we can specify some options when we call `ivregress`. Add the `, firststage` option and compare the output to the output from your manual estimation of the first stage.

14. You can also use `estat firststage` to instead just report summary statistics for the first stage without the coefficient estimates. Note how much of the output from `estat firststage` is the same as the summary statistics we obtain from any regression. The key statistic here is the F-statistic. Recall that in a regression with one regressor, the F-statistic is simply the square of the t-statistic for the coefficient. A rule-of-thumb is that an F-statistic greater than 10 is required for the instrument to be considered strong.

15. The bottom half of the `estat firststage` output is something new. It lists critical values (in F-statistic terms) above which we can reject a null hypothesis of having a weak instrument(s). Does the F-statistic exceed these critical values?

Tests of endogeneity

16. While we cannot test the exclusion restriction (that the instrument has no direct effect on the outcome), we can test whether the seemingly endogenous regressor is in fact endogenous. We do this by comparing the OLS estimates of the regressor's effects to those estimated by 2SLS. Or, to put it another way, we see whether the portion of X not explained by the instrument has any independent effect on Y. Stata provides this test — called the Durbin-Wu-Hausman Test — using `estat endogenous`. We are interested in the last line of output, which provides an F-statistic and associated p-value for a test of the null hypothesis that the included covariates in the second stage are exogenous. Rejecting the null would indicate that the include variable(s) are endogenous and should be instrument. We have one included variable (`logrprice`). Does it seem to be exogenous?

17. To get a feel for what's happening in the DWH test, let's calculate it manually. First, rerun the first-stage regression. Then save the residuals from the firststage as a new variable `resid`. Now regress `logpacks` on `logrprice` and `resid`. If you run `test resid`, the F-statistic and p-value should match those of from the output of `estat endogenous`.

Covariates in IV estimation

18. We don't simply need to estimate a model with one endogenous regressor and one instrument. We can also include additional covariates in our model. Perhaps there were other variables that we think confound the relationship between `logrprice` and `logpacks` and that we were able to observe directly. We can include these in our model. For example, maybe we think that state income `logincome` has an effect on smoking. Estimate a two-stage least squares model that includes this as an additional covariate. Does it seem to be influential?

19. To get a handle on what happens when we include additional covariates in our model, run the following:
```
quietly reg logrprice tdiff logincome
predict pricefit2
reg logpacks pricefit2 logincome
```
Here we manually calculate the two-stage least squares estimates, as we did above. Note how we include `logincome` in both the first- and second-equations. The reason for this that we when we regress X on the instrument and all of the other covariates, the residuals from that first-stage model are — by definition — independent of the instrument and the covariates. Thus when we estimate the second-stage equation using the instrument $\hat{X}$, we leave some of the original variation from X in the second-stage error term, but that variation is independent of $\hat{X}$ and the other covariates.

Multiple instruments

20. What if we happen to be in the rare situation where we have multiple instruments? Well, they allow us to additional conduct tests of "overidentifying restrictions." (You should see Wooldridge for statistical details here.) This is a test of the null hypothesis that all instruments are valid. Rejecting it suggests that at least one of the instruments is invalid. Try running a 2SLS estimation using `population` as an additional instrument. Then use the `estat overid` to test the validity of the combined instruments.

21. In practice, we will rarely have one instrument let alone two or more. As such, we are restricted to a world where there is only one included variable that is affected by unobserved confounding. Having two instruments would mean we could estimate a model with two endogenous regressors if both instruments are credibly exogenous for both of these variables. Feel free to estimate such a model.

Model fit and specification

22. Remember that 2SLS is just like OLS, it is a linear function of covariates. For this reason, all the usual rules about goodness-of-fit, model specification, and linearity of the CEF apply. Using tools we learned for OLS, test whether the outcome is a linear function of the covariates.

23. As in OLS, we can also estimate heteroskedasticity-consistent standard errors using the `, vce(robust)` (or its shortcut, `, r`) option. See whether this is appropriate using a Breusch-Pagan test (but note that this is not implemented as a postestimation command after `ivregress`; instead estimate the model manually and then use the usual `hettest` command).

24. This dataset is relatively sparse in terms of variables, but at this point you could also test empirically whether there are other currently excluded variables that perhaps should be in the model.

TABLE 6.2.1
OLS and fuzzy RD estimates of the effect of class size on fifth-grade math scores

	OLS			2SLS				
				Full Sample		Discontinuity Samples		
						±5		±3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean score (SD)		67.3 (9.6)			67.3 (9.6)		67.0 (10.2)	67.0 (10.6)
Regressors								
Class size	.322 (.039)	.076 (.036)	.019 (.044)	−.230 (.092)	−.261 (.113)	−.185 (.151)	−.443 (.236)	−.270 (.281)
Percent disadvantaged		−.340 (.018)	−.332 (.018)	−.350 (.019)	−.350 (.019)	−.459 (.049)	−.435 (.049)	
Enrollment			.017 (.009)	.041 (.012)	.062 (.037)		.079 (.036)	
Enrollment squared/100					−.010 (.016)			
Segment 1 (enrollment 38–43)								−12.6 (3.80)
Segment 2 (enrollment 78–83)								−2.89 (2.41)
R^2	.048	.249	.252					
Number of classes		2,018			2,018		471	302

Notes: Adapted from Angrist and Lavy (1999). The table reports estimates of equation (6.2.6) in the text using class averages. Standard errors, reported in parentheses, are corrected for within-school correlation.

Fuzzy regression discontinuity

25. Load the `AngristPischke621.dta` dataset. This contains the data reported in Angrist and Pischke Table 6.2.1 (p.266).

26. One of the most compelling applications of IV estimation is in the presence of regression discontinuities. A discontinuity is (a potentially striking) change in the value of a treatment variable across levels of an instrument. A nice example of this that we've encountered is Maimonides' Rule and its effect on class sizes in Israel. Because class sizes are reduced dramatically when total school enrollment approaches any multiple of 40 students, class sizes at schools just above and just below those thresholds are as-if randomly assigned. That is to say, Maimonides' Rule creates an experiment in schools of around 40, 80, 120, 160, etc. students.

27. Angrist and Lavy (1999) were interested in the effect of class size (`classize`) on student achievement (measured by math [`avgmath`] and reading [`avgverb`] test scores). Draw a simple causal graph representing this relationship, highlighting the included variables shown in Table 6.2.1, unmeasured confounding of the relationship between class size and test scores, and the instrumental variable (school enrollment).

28. Focusing on the math outcome (`avgmath`), estimate the OLS models (1,2,3) from Table 6.2.1. The relevant variables from the table are:

    - `classize` : Class size
    - `tipuach` : Percent disadvantaged
    - `c_size` : Enrollment (School cohort size)

- `c_size2` : $Enrollment^2/100$

Confirm that your coefficient estimates correspond with those reported in the table. Note: The SEs will differ from the published result, due to some procedures Angrist and Lavy used to address clustering of students within schools, which are unimportant for our purposes.

29. The instrument here is school enrollment, but recall that there are multiple discontinuities, so the endogenous variable (class size) is not a linear function of the instrument. You can (sort of) see this in a scatterplot (like Figure 6.2.1, panel A): `twoway scatter classize c_size`

30. To address this, Angrist and Lavy calculate the class size expected by Maimonides' Rule as a function of school enrollment. This is stored as `func1`. Use a scatterplot to check for the linear of the relationship between this transformed version of the enrollment instrument and class size.

31. Use methods from earlier to assess instrument relevance.

32. Estimate the "Full sample" 2SLS models (4,5), instrumenting `classize` with `func1`.

33. Perform relevant postestimation commands.

34. One concern in a regression discontinuity approach is that the treatment (here class size) is only as if random right at the point of the discontinuity (here discontinuities). In short, the previous results use the full sample data to estimate the effect of class size on test scores even though schools with, e.g., 20 students or 105 students are not anywhere near the Maimonides' Rule thresholds (40, 80, 120, etc.). As a result, it is common in regression discontinuity designs to focus only on cases that are "near" the discontinuity threshold. The dataset contains a variable `disc` that is 1 if a student was in a school that had fell within 5 students of a threshold (e.g., between 36 and 45). To get a sense of this subsample indicator variable, use `tab` and `twoway scatter disc c_size` to see which students are included in this causally more credible subsample.

35. Using the `if` statement to subset the data during estimation, estimate the "discontinuity sample" models (6,7).

36. The last column of Table 6.2.1 uses an even narrower subset of students (those within 3 students of discontinuity threshold). Generate a new variable `disc3` that serves as an indicator for these students. Additionally, you need to create indicators for which group the student is in. With those three variables, run the 2SLS estimate shown in the last column. Don't worry if your results don't match exactly but you should be able to match them to within about 0.1.

37. You can repeat the above analyses using `avgverb` as the outcome.