R Applications for Instrumental Variable Estimation and 2SLS

Introductory Remarks

This file builds on the previous R Activities. If you have not yet gone through those files, please make sure to do so before tackling this R Activity.

Wage Equation Set-Up

Similarly to the lecture, we are concerned with estimating the following wage equation

$$log(wage_i) = \alpha_0 + \alpha_1 educ_i + \epsilon_i$$

In words, we want to look at the effects of education on log wages. To do this we will use the "CARD" dataset, available in the "wooldridge" package in R. After installing and loading the package, type

```
data("card") # load data
```

As you can see, the dataset consists of 3010 observations and 34 variables. From the lecture we know that *educ* above is an endogeneous variable. As such, running a simple OLS regression on the above regression will yield a biased estimate. We will therefore need (at least) one instrument.

2SLS/IV

Using One Instrument

We begin our R estimation of the above equation with one instrument, namely *nearc*4, a dummy variable indicating proximity to a four-year college, as our instrument. In order to be a valid instrument, we need this distance measure (a) to be correlated with education (first stage: the instrument and the endogeneous regressor are correlated), and (b) not to affect log wages other than through education (exogeneity: the instrument and the error are uncorrelated). From the lecture we know that we cannot test our exogeneity assumption (you can try to think for yourself whether you think that it is satisfied in this case), but we can test the first stage.

Luckily the first stage is just an OLS regression of *educ* on *nearc*4, which we already know how to do in R.

```
card_FS1 <- lm(educ ~ nearc4, data=card) # first stage reg
coeftest(card_FS1, vcov = vcovHC(card_FS1, type = "HC1")) # heterosk. SEs
```

You are by now familiar with the first line of code. The second line, instead of using the "summary" command to display the results uses the "coeftest" command in order to be able to report heteroskedastic robust standard errors. (Make sure you have installed the packages "lmtest" and "sandwich" for this.) The first stage requirement is that *educ* and *nearc*4 are strongly correlated. We can test this by looking at whether the coefficient on *nearc*4 (0.829 in our regression) is statistically significantly different from zero. Given the t-statistic of 7.77 (or the p-value that's essentially zero), we conclude that this coefficient is statistically significantly different from zero. Thus, our first stage requirement is given.

We can therefore now try to run the IV regression we are interested in in R. To do this, we use a command called "ivreg" from the AER package, i.e. make sure you've loaded and installed this package. To run this regression, we type

```
card_IV1 <- ivreg(lwage ~ educ | nearc4, data=card) # IV reg w/ one instrument
coeftest(card_IV1, vcov = vcovHC(card_IV1, type = "HC1")) # heterosk. SEs
```

The first line of code here runs the actual IV regression. Specifically, we start by calling the function "ivreg." The first argument is then the relationship we are interested in, which in our case is the regression of log wages on education. The second argument comes after "|", which is akin to us telling R that whatever variable comes after "|" should be used as an instrument. Lastly, as always, we have to tell R what dataset we're using. The second line of code displays heteroskedasticity robust standard errors again.

The coefficient we find is .188 with a standard error of (.026). Clearly, it is statistically significant (even at the 1% level). The interpretation is as follows: a one year increase in education iscreases log wages by 18.8%. Alternatively, and more elegantly, this regression tells us that the return to education is 18.8%.

Adding Controls

The above is likely to be a bit too simplistic. Indeed, we may want to add some controls to our regression in order to try to capture any possible confounders we can think of. To add controls into an IV regression in R, we can type

```
card_IV1C <- ivreg(lwage ~ educ + exper + expersq + black +
                   smsa + south + smsa66 + reg662 + reg663 + reg664 +
                   reg665 + reg666 + reg667 + reg668 + reg669 |
                   nearc4 + exper + expersq + black +
                   smsa + south + smsa66 + reg662 + reg663 + reg664 +
                   reg665 + reg666 + reg667 + reg668 + reg669, data=card)
coeftest(card_IV1C, vcov = vcovHC(card_IV1C, type = "HC1")) # heterosk. SEs
```

As you can see, the procedure is analogue to the above, with the exception that we have added the controls we want to add to our regression after the endogenous variable as well as after the instrument. It is crucial that you add these controls also in the first stage and therefore they need to be after the "|" as well. The controls we include are experience, experience squared, a dummy if the individual is African American, as well as a range of region and geography dummies for the USA. As you can see, the resulting coefficient on *educ* is .1315, which indicates a return of 13.2% to education. It is slightly lower than above when we ran the model without controls.

Using Two Instruments

From the lecture you know that you can also use two instruments instead of just one as we did above. Indeed, if we wanted to we could also use five instruments (if we can come up with them). The crucial point is that we need at least as many instruments as we have endogenous regressions. In our case, this means that since we have one endogeneous regressor (education), we need at least one instrument. However, this dataset provides us with a second possible instrument, namely *nearc*2, a dummy variable indicating proximity to a two-year college. Similarly as above, we can test the first stage but not the exogeneity assumption.

To test the first stage assumption, i.e. that *educ* is correlated with *nearc*4 and *nearc*2, we run the following OLS regression

```
card_FS2 <- lm(educ ~ nearc4 + nearc2 + exper + expersq + black +
                 smsa + south + smsa66 + reg662 + reg663 + reg664 +
                 reg665 + reg666 + reg667 + reg668 + reg669, data=card)
coeftest(card_FS2, vcov = vcovHC(card_FS2, type = "HC1")) # heterosk. SEs
```

Notice that this is analogous to the above, except that we now regress education on both our instruments as well as all the controls. To test whether the first stage is satisfied in this case, we need to test the hypothesis whether the coefficients on *nearc*2 and *nearc*4 are jointly significant. For this, we need an F-test. To do this, we can type the following

```
linearHypothesis(card_FS2, c("nearc4", "nearc2"),
                 c(0, 0), vcov=vcovHC(card_FS2, type="HC1")) # F-test
```

As you can see, the resulting F-statistic is 8.319 and is highly statistically significantly different from zero. That being said, the "magic" number we look for in fist stage F-statistics is ofter 10, i.e. we often aim to have an F-statistic above 10 in a first stage. For our purposes, we ignore this issue for now and proceed nonetheless.

To run the 2SLS regression with both instruments and the controls, we can now type

```
card_2SLS <- ivreg(lwage ~ educ + exper + expersq + black +
                   smsa + south + smsa66 + reg662 + reg663 + reg664 +
                   reg665 + reg666 + reg667 + reg668 + reg669 |
                   nearc4 + nearc2 + exper + expersq + black +
                   smsa + south + smsa66 + reg662 + reg663 + reg664 +
                   reg665 + reg666 + reg667 + reg668 + reg669, data=card)
coeftest(card_2SLS, vcov = vcovHC(card_2SLS, type = "HC1")) # heterosk. SEs
```

As you can see, this just combines everything we have done above into the "ivreg" command. The only addition is that the second instrument is just added after the "|" and the first instrument. The coefficient on *educ* is 0.157 and indicates that the returns to education are 15.7%. Notice that this estimate is between the first and second estimates we got above. Reassuringly, all three estimates are similar in magnitude.

Testing for Endogeneity

From the lecture we know that to test for endogeneity, we can follow a simple two step procedure. First, we estimate the first stage regression and save the residuals. Second, we estimate the our equation of interest but add these residuals into the regression. If the coefficient on the residuals is significantly different from zero, we have an endogeneity problem.

To do this in R, consider again the original case where we neglect all controls and only use *nearc*4 as an instrument. The first step of this procedure is therefore to run the first stage equation from above again, i.e.

```
card_FS1 <- lm(educ ~ nearc4, data=card) # first stage reg
card$v <- card_FS1$residuals # save residuals
```

The second line then just saves these residuals as an additional variable in our dataset. Intuitively, these residuals are the part of education that is not explained by *nearc*4, our instrument.

The second part of this procedure then runs our original wage equation but includes these residuals as well. In other words, we regress log wages on education and the residuals from above, i.e. we type

```
end_test <- lm(lwage ~ educ + v, data=card)
coeftest(end_test, vcov = vcovHC(end_test, type = "HC1")) # heterosk. SEs
```

The output of this regression suggests that the coefficient on v is $-.139$ with a p-value of essentially zero. In other words, the coefficient is highly statistically significantly different from zero. As such, we conclude that education is indeed an endogenous variable.

Note that in this example it was obvious that education is an endogenous variable. Indeed, often times researchers do not test for endogeneity because it is very obvious that endogeneity will be an issue. That being said, it is useful to know this procedure and to understand how to implement it in R.