Recitation 8 Notes: Instrumental Variables

Matthew Alampay Davis

December 5, 2021

Instrumental variables regressions

Let's use the following dataset:

```
daron <- read.dta13('maketable4.dta')
```

This dataset was used in a famous study seeking to further the understanding of why countries farther from the equator seem to enjoy higher economic growth. The authors attribute the difference to a pattern of European colonization which was different in different climates. Areas in which Europeans had more resistance to the local diseases were colonized such that governmental institutions came to resemble those in Europe. Areas where Europeans had less resistance to the local diseases were colonized such that governmental institutions were created to extract wealth from the colony and its local inhabitants.

The authors have assessments of current government institutions in many countries, and of the government institutions that existed in the past, and data on GDP. They would like to show that government institutions are a large factor in determining economic well-being. They also would like to show that due to the extraordinarily high persistence of culture, current institutions seem to have their roots in the distant past, particularly the in the experience of European colonization. Economists sometimes use the word strong to indicate democratic-type institutions and the word weak to indicate institutions that do not offer much protection from rights violations and expropriation.

But the authors worry about endogeneity. As an example, the colonists' wealth could feed back into their choice of government institutions. If that is true, then it is some aspect of colonization other than disease resistance that may have caused the path-dependent outcome of institutions. Or it may have been the religion or some other aspect of the culture of particular European groups that caused different paths for institutions.

Instrumental variables and two-stage least squares provide a mechanism to examine only the input of European disease resistance during the colonization period on the formulation of governmental institutions. The governmental institutions index is regressed upon the European mortality rate during colonization, and control variables. The predicted governmental institutions strength can then be interpreted as the portion of institutional strength that can be attributed to the European disease resistance. The outcome of these regressions is quite interesting on its own, and the authors always report both stages of their two-stage process.

The predicted institutional strength is then regressed on GDP to show that past European disease resistance is a major factor in explaining the current arrangement of GDP around the world, and the linkage between the past experience and the current situation is governmental institutions.

Short version of above:

Authors want to measure the effect of institution type on modern economic growth. They were worried that institution type is endogeneous to growth. They argue that modern institution type is likely related to the

choice of institutions that European colonists chose to put in place in a given area and that they likely chose "good" institutions in areas where they had disease resistance and chose "extractive" institutions (bad for growth) in areas where they did not have strong disease resistance. So per capita mortality is a candidate for an instrument that can enable IV estimation of the effect of institutions on growth.

Instrumental variables regressions

To do 2SLS, let's use "iv_robust", which like "lm_robust" comes from the *estimatr* package and takes in an "se_type" argument as well. Important to note here that for IV regressions, we'll want to indicate "se_type = 'HC0' " for our choice of heteroskedasticity-robust standard errors. Remember we found that "se_type = 'stata' " is the same as "se_type = 'HC1' ". BUT for some reason, Stata defaults to a different formula ('HC0') when doing IV regressions. I don't know why (everyone should be using 'HC3' anyway to be honest), but we'll want to match Stata solutions so please remember this.

So let's run our first 2SLS regression where

- *logpgp95* is the outcome variable, log 1995 GDP
- *avexpr* is the endogenous variable, the average (contemporary) protection against expropriation risk (consider this a measure of institution quality)
- *logem4* is the instrumental variable, the log of the colonial settler mortality rate (consider this a measure of colonist disease resistance of a particular area)
- *lat_abst* is a measure of the latitude which we use as a control variable

So let's run the IV regression without the control variable: one endogenous variable and one instrument:

```
mod.1 <- iv_robust(logpgp95 ~ avexpr | logem4,
                   data = daron,
                   se_type = 'HC0') # Note the choice of standard errors is different!
summary(mod.1)
```

```
##
## Call:
## iv_robust(formula = logpgp95 ~ avexpr | logem4, data = daron,
##     se_type = "HC0")
##
## Standard error type:  HC0
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## (Intercept)   2.3702     0.9283   2.553 1.292e-02   0.5178    4.222 68
## avexpr        0.8684     0.1365   6.361 1.972e-08   0.5960    1.141 68
##
## Multiple R-squared:  0.3045 ,    Adjusted R-squared:  0.2942
## F-statistic: 40.46 on 1 and 68 DF,  p-value: 1.972e-08
```

Note the use of the '|' in the formula. On the left side of the | are the regressors as given in the model with the endogenous variables and control variables (no controls in this case). On the right side are all the regressors but replacing the endogenous variables with instruments (in this case, we had one endogenous regressor which is replaced with one instrument). Note that estimation is only possible if there are at least as many regressors on the right as on the left, i.e., the model is just-identified or overidentified (exogenous regressors appear on both sides).

To make this more concrete, consider a second model where *lat_abst* is used as a control variable (or an additional exogenous regressor). This means we only want to instrument for *avexpr* not *lat_abst*.

```
mod.2 <- iv_robust(logpgp95 ~ lat_abst + avexpr | lat_abst + logem4, data = daron, se_type = 'HC0')
summary(mod.2)
```

```
##
## Call:
## iv_robust(formula = logpgp95 ~ lat_abst + avexpr | lat_abst +
##     logem4, data = daron, se_type = "HC0")
##
## Standard error type:  HC0
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## (Intercept)    1.652      1.515   1.090 0.2794969  -1.3722    4.676 67
## lat_abst      -1.784      1.530  -1.166 0.2476524  -4.8383    1.270 67
## avexpr         1.029      0.263   3.913 0.0002155   0.5042    1.554 67
##
## Multiple R-squared:  0.06701 ,   Adjusted R-squared:  0.03915
## F-statistic: 19.39 on 2 and 67 DF,  p-value: 2.276e-07
```

So we included *lat_abst* on both sides of the bar "|" in the formula argument. If we hadn't included it on the right side, then the model would assume *lat_abst* was another endogenous variable and it'd have more endogenous variables than instruments, which would prevent the regression from running.

And we can do the same with even more controls. Say we include all the continent dummies as control variables:

```
daron %<>% mutate(other = shortnam %in% c('AUS', 'NZL', 'MLT')) # Countries not in Africa or Asia
mod.3 <- iv_robust(logpgp95 ~ avexpr + africa + asia + other + lat_abst |
                     logem4 + africa + asia + other + lat_abst,
                   data = daron,
                   se_type = 'HC0')
summary(mod.3)
```

```
##
## Call:
## iv_robust(formula = logpgp95 ~ avexpr + africa + asia + other +
##     lat_abst | logem4 + africa + asia + other + lat_abst, data = daron,
##     se_type = "HC0")
##
## Standard error type:  HC0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)   2.7569     2.1253  1.2971 0.199237  -1.4890  7.00268 64
## avexpr        0.9162     0.3585  2.5554 0.012993   0.1999  1.63246 64
## africa       -0.5820     0.3041 -1.9140 0.060090  -1.1895  0.02545 64
## asia         -0.9276     0.3444 -2.6935 0.009017  -1.6156 -0.23962 64
## otherTRUE    -0.5009     0.7050 -0.7104 0.480019  -1.9093  0.90757 64
## lat_abst     -1.5287     1.5771 -0.9693 0.336029  -4.6792  1.62184 64
##
## Multiple R-squared:  0.3648 ,    Adjusted R-squared:  0.3152
## F-statistic: 10.27 on 5 and 64 DF,  p-value: 2.944e-07
```

Overidentification tests

If we wanted to test for weak instruments or do an overidentification test, we'd add an argument to "iv_robust" called "diagnostics":

```
model.tests <- iv_robust(logpgp95 ~ avexpr | lat_abst + logem4,
                         data = daron, se_type = 'HC0',
                         diagnostics = T) # New argument
summary(model.tests)
```

```
##
## Call:
## iv_robust(formula = logpgp95 ~ avexpr | lat_abst + logem4, data = daron,
##     se_type = "HC0", diagnostics = T)
##
## Standard error type:  HC0
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## (Intercept)   2.8406     0.7255   3.916 2.114e-04   1.3930    4.288 68
## avexpr        0.7976     0.1051   7.586 1.239e-10   0.5878    1.007 68
##
## Multiple R-squared:  0.4009 ,    Adjusted R-squared:  0.3921
## F-statistic: 57.54 on 1 and 68 DF,  p-value: 1.239e-10
##
## Diagnostics:
##                   numdf dendf  value  p.value
## Weak instruments      2    67 24.899 8.21e-09 ***
## Wu-Hausman            1    67 19.985 3.09e-05 ***
## Overidentifying       1    NA  2.335    0.126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get the same regression results but now have three new test results under diagnostics. The Weak instruments test gives us the F-statistic under the "value" column. The Wu-Hausman test returns the test for the endogeneity of the endogenous variables (we do not cover this in class). The overidentifying test gives the results for the test of the hypothesis that all instruments are exogenous. We can also run the following:

```
model.tests$diagnostic_first_stage_fstatistic
```

```
##        value        nomdf        dendf      p.value
## 2.489862e+01 2.000000e+00 6.700000e+01 8.213216e-09
```

```
model.tests$diagnostic_overid_test
```

```
##    value       df  p.value
## 2.335455 1.000000 0.126458
```

Empirical Exercise 12.1

How does fertility affect labor supply? That is, how much does a woman's labor supply fall when she has an additional child? In this exercise, you will estimate this effect using data for married women from the 1980 U.S. Census.

The data set contains information on married women aged 21–35 with two or more children:

```
fertility <- read.dta13('fertility.dta')
head(fertility)
```

```
##   morekids boy1st boy2nd samesex agem1 black hispan othrace weeksm1
## 1        0      1      0       0    27     0      0       0       0
## 2        0      0      1       0    30     0      0       0      30
## 3        0      1      0       0    27     0      0       0       0
## 4        0      1      0       0    35     1      0       0       0
## 5        0      0      0       1    30     0      0       0      22
## 6        0      1      0       0    26     0      0       0      40
```

Part a: Regress weeks worked on the indicator variable morekids, using OLS. On average, do women with more than two children work less than women with two children? How much less?

```
mod.a <- lm_robust(weeksm1 ~ morekids, data = fertility, se_type = 'stata')
summary(mod.a)
```

```
##
## Call:
## lm_robust(formula = weeksm1 ~ morekids, data = fertility, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper     DF
## (Intercept)   21.068    0.05607  375.76        0   20.959   21.178 254652
## morekids      -5.387    0.08715  -61.81        0   -5.558   -5.216 254652
##
## Multiple R-squared:  0.01431 ,   Adjusted R-squared:  0.0143
## F-statistic:  3821 on 1 and 254652 DF,  p-value: < 2.2e-16
```

The coefficient is -5.387, which indicates that women with more than 2 children work 5.387 fewer weeks per year than women with 2 or fewer children.

Part b: Explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (morekids) on labor supply (weeksworked).

Both fertility and weeks worked are choice variables. A women with a positive labor supply regression error (a women who works more than average) may also be a woman who is less likely to have an additional child. This would imply that Morekids is positively correlated with the regression error, so that the OLS estimator on the coefficient of morekids is positively biased.

Part c: The data set contains the variable samesex, which is equal to 1 if the first two children are of the same sex (boy–boy or girl–girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child? Is the effect large? Is it statistically significant?

```
mod.c <- lm_robust(morekids ~ samesex, data = fertility, se_type = 'stata')
summary(mod.c)
```

```
##
## Call:
## lm_robust(formula = morekids ~ samesex, data = fertility, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper     DF
## (Intercept)  0.34642   0.001341  258.34  0.000e+00  0.34380  0.34905 254652
## samesex      0.06753   0.001919   35.19 1.388e-270  0.06376  0.07129 254652
##
## Multiple R-squared:  0.004835 ,  Adjusted R-squared:  0.004831
## F-statistic:  1238 on 1 and 254652 DF,  p-value: < 2.2e-16
```

Couples with samesex = 1 are 6.6% more likely to have an additional child that couples with samesex = 0. The effect is highly significant.

Part d: Explain why samesex is a valid instrument for the IV regression of weeksworked on morekids.

Samesex is random and is unrelated to any of the other variables in the model including the error term in the labor supply equation. Thus, the instrument is exogenous. From (c), the first stage F-statistic is large (F = 1238) so the instrument is relevant. Together, these imply that samesex is a valid instrument.

Part e: Is samesex a weak instrument?

No, for the reason in (d)

Part f: Estimate the IV regression of weeksworked on morekids, using samesex as an instrument. How large is the fertility effect on labor supply?

```
mod.f <- iv_robust(weeksm1 ~ morekids | samesex, data = fertility, se_type = 'HC0')
summary(mod.f)
```

```
##
## Call:
## iv_robust(formula = weeksm1 ~ morekids | samesex, data = fertility,
##     se_type = "HC0")
##
```

```
## Standard error type:  HC0
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper     DF
## (Intercept)   21.421     0.4872  43.963 0.000e+00   20.466   22.376 254652
## morekids      -6.314     1.2747  -4.953 7.307e-07   -8.812   -3.815 254652
##
## Multiple R-squared:  0.01388 ,   Adjusted R-squared:  0.01388
## F-statistic: 24.53 on 1 and 254652 DF,  p-value: 7.307e-07
```

The estimated coefficient is -6.314.

Part g: Do the results change when you include the variables agem1, black, hispan, and othrace in the labor supply regression (treating these variable as exogenous)? Explain why or why not.

```
mod.g <- iv_robust(weeksm1 ~ morekids + agem1 + black + hispan + othrace |
                     samesex + agem1 + black + hispan + othrace,
                   data = fertility, se_type = 'HC0')
summary(mod.g)
```

```
##
## Call:
## iv_robust(formula = weeksm1 ~ morekids + agem1 + black + hispan +
##     othrace | samesex + agem1 + black + hispan + othrace, data = fertility,
##     se_type = "HC0")
##
## Standard error type:  HC0
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper     DF
## (Intercept)  -4.7919    0.38979  -12.29  1.003e-34  -5.5559  -4.0279 254648
## morekids     -5.8211    1.24639   -4.67  3.008e-06  -8.2639  -3.3782 254648
## agem1         0.8316    0.02264   36.73 1.422e-294   0.7872   0.8760 254648
## black        11.6233    0.23180   50.14  0.000e+00  11.1690  12.0776 254648
## hispan        0.4042    0.26080    1.55  1.212e-01  -0.1070   0.9153 254648
## othrace       2.1310    0.21099   10.10  5.580e-24   1.7174   2.5445 254648
##
## Multiple R-squared:  0.04368 ,   Adjusted R-squared:  0.04366
## F-statistic:  1391 on 5 and 254648 DF,  p-value: < 2.2e-16
```

The results do not change in an important way. The reason is that samesex is unrelated to agem1, black, hispan, othrace, so that there is no omitted variable bias in IV regression in (2).