Instrumental Variables

Jeffrey Grove

May 17, 2017

Question 1

(a)

```
reg_1 <- tidy(lm(InternalConflict ~ LaggedGDPGrowth, data = data))

reg_1
```

```
##              term    estimate  std.error  statistic      p.value
## 1     (Intercept)  0.26737746 0.01631415 16.3892984 9.586655e-52
## 2 LaggedGDPGrowth -0.08206485 0.22485213 -0.3649725 7.152360e-01
```

The results of the bivariate OLS do no demonstrate significance at the alpha equals 0.05 level, thus we do not reject the null hypothesis. ###(b)

```
reg_2 <- tidy(lm(InternalConflict ~ LaggedGDPGrowth + InitialGDP + Democracy + Mountains + EthnicFrac +

reg_2
```

```
##              term     estimate    std.error  statistic      p.value
## 1     (Intercept)  0.070355529 0.0731012386  0.9624396 3.361449e-01
## 2 LaggedGDPGrowth -0.108797693 0.2200998529 -0.4943106 6.212343e-01
## 3      InitialGDP -0.056909063 0.0182258230 -3.1224413 1.863801e-03
## 4       Democracy  0.001224162 0.0028894131  0.4236714 6.719292e-01
## 5       Mountains  0.003865434 0.0009526937  4.0573730 5.493161e-05
## 6      EthnicFrac  0.324793069 0.0918181328  3.5373521 4.295018e-04
## 7   ReligiousFrac  0.010516152 0.0958907296  0.1096681 9.127025e-01
```

These results do not establish a causal relationship between the economy and civil conflict, as the p value is still greater than alpha at a 0.05 level. ###(c)

```
itest <- tidy(lm(LaggedGDPGrowth ~ LaggedRainfallGrowth + InitialGDP + Democracy + Mountains + EthnicFra

itest
```

```
##                   term      estimate    std.error  statistic      p.value
## 1          (Intercept) -0.0058040646 0.0121558449 -0.4774711 0.6331684550
## 2 LaggedRainfallGrowth  0.0439769947 0.0128127339  3.4322881 0.0006319632
## 3           InitialGDP -0.0008055799 0.0030286889 -0.2659830 0.7903267540
## 4            Democracy  0.0005373436 0.0004797716  1.1199986 0.2630797024
## 5            Mountains  0.0001086811 0.0001582464  0.6867842 0.4924350511
## 6           EthnicFrac  0.0031602813 0.0152584101  0.2071173 0.8359755120
## 7        ReligiousFrac -0.0017176029 0.0159357476 -0.1077830 0.9141971916
```

The two conditions required fora good instrument are the inclusion condition - the instrument must explain x - and the exclusion restriction - the instrument must not explain Y. We can test for the first using a standard linear regression as done above. We find that the instrument rainfall explains growth at the alpha equals 0.05 level. However, the only way to justify the second condition is through theoretical explanation, we can not use a statistical test to establish its veracity.

(d)

Instrumenting for GDP with rain could explain the causal effect as rainfall would help explain overall economic growth, especially in agrarian economies. Yet, importantly, it is unlikely that rainfall would have a strong effect on whether conflict itself occurs, except through economic growth itself. Acemoglu and Robinson test the hypothesis that rainfall would be correlated to the destruction of infrastructure, particularly roads, but find no particular evidence of this alternate causal path.

(e)

```
ireg <- summary(ivreg(InternalConflict ~ LaggedGDPGrowth + InitialGDP + Democracy + Mountains + EthnicFr

ireg
```

```
##
## Call:
## ivreg(formula = InternalConflict ~ LaggedGDPGrowth + InitialGDP +
##     Democracy + Mountains + EthnicFrac + ReligiousFrac | LaggedRainfallGrowth +
##     InitialGDP + Democracy + Mountains + EthnicFrac + ReligiousFrac,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1693 -0.3106 -0.1897  0.4203  2.0093
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.062506   0.077268   0.809 0.418802
## LaggedGDPGrowth -2.063153   1.845106  -1.118 0.263857
## InitialGDP      -0.058080   0.019209  -3.024 0.002584 **
## Democracy        0.002361   0.003221   0.733 0.463785
## Mountains        0.004069   0.001020   3.988 7.34e-05 ***
## EthnicFrac       0.328851   0.096686   3.401 0.000707 ***
## ReligiousFrac    0.004724   0.101042   0.047 0.962721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.456 on 736 degrees of freedom
## Multiple R-Squared: -0.05059,    Adjusted R-squared: -0.05916
## Wald test: 6.133 on 6 and 736 DF,  p-value: 2.748e-06
```

From this regression, we find that Lagged GDP Growth is still not a statistically significant explanation at the alpha equals 0.05 level, and thus do not reject the null hypothesis.

(f)

```
summary(ivreg(InternalConflict ~ LaggedGDPGrowth + InitialGDP + Democracy + Mountains + EthnicFrac + Rel
```

```
##
## Call:
## ivreg(formula = InternalConflict ~ LaggedGDPGrowth + InitialGDP +
##     Democracy + Mountains + EthnicFrac + ReligiousFrac + country_name |
##     LaggedRainfallGrowth + InitialGDP + Democracy + Mountains + 
##         EthnicFrac + ReligiousFrac + country_name, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.60872 -0.18282 -0.01501  0.13649  1.92662 
## 
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                               0.459156   1.863122   0.246
## LaggedGDPGrowth                          -2.853380   1.535631  -1.858
## InitialGDP                               -0.476826   0.792632  -0.602
## Democracy                                 0.001065   0.003276   0.325
## Mountains                                 0.092267   0.020715   4.454
## EthnicFrac                               -0.177308   0.960665  -0.185
## ReligiousFrac                             0.122734   1.081919   0.113
## country_nameBenin                         0.102932   0.151373   0.680
## country_nameBotswana                      0.459522   0.379597   1.211
## country_nameBurkina Faso                  0.001897   0.456368   0.004
## country_nameBurundi                      -6.759628   1.224003  -5.523
## country_nameCameroon                     -1.449536   0.539017  -2.689
## country_nameCentral African Republic     -0.527030   0.202258  -2.606
## country_nameChad                         -0.049368   0.121434  -0.407
## country_nameCongo                         0.594180   0.520803   1.141
## country_nameDjibouti                      0.108725   0.362860   0.300
## country_nameEthiopia                     -5.942561   1.157047  -5.136
## country_nameGabon                         1.894540   2.969182   0.638
## country_nameGambia                        0.146595   0.339041   0.432
## country_nameGhana                         0.160302   0.216191   0.741
## country_nameGuinea                       -0.165534   0.412092  -0.402
## country_nameGuinea-Bissau                -0.012907   0.290915  -0.044
## country_nameIvory Coast                   0.570719   1.158194   0.493
## country_nameKenya                        -2.322296   0.572570  -4.056
## country_nameLesotho                      -7.505690   1.488785  -5.041
## country_nameLiberia                       0.260105   0.270259   0.962
## country_nameMadagascar                   -3.146483   0.563717  -5.582
## country_nameMalawi                       -0.962630   0.304634  -3.160
## country_nameMali                         -0.011005   0.632748  -0.017
## country_nameMauritania                    0.051883   0.865955   0.060
## country_nameMozambique                    0.502944   0.170638   2.947
## country_nameNamibia                       0.191278   1.379937   0.139
## country_nameNiger                         0.029188   0.386649   0.075
## country_nameNigeria                       0.016553   0.535843   0.031
## country_nameRwanda                       -6.447391   1.211757  -5.321
## country_nameSenegal                       0.581121   0.314757   1.846
## country_nameSierra Leone                  0.384331   0.279397   1.376
## country_nameSomalia                      -0.202700   0.986690  -0.205
## country_nameSouth Africa                  1.195777   2.036813   0.587
## country_nameSudan                         0.352135   0.189502   1.858
## country_nameSwaziland                    -0.336431   1.607692  -0.209
## country_nameTanzania, United Republic of -2.107378   0.320915  -6.567
## country_nameTogo                          0.014390   0.326231   0.044
##                                          Pr(>|t|)    
## (Intercept)                               0.80541    
## LaggedGDPGrowth                           0.06357 .  
## InitialGDP                                0.54765    
## Democracy                                 0.74518    
## Mountains                                9.81e-06 ***
## EthnicFrac                                0.85362    
## ReligiousFrac                             0.90971    
## country_nameBenin                         0.49674    
## country_nameBotswana                      0.22648    
## country_nameBurkina Faso                  0.99668    
## country_nameBurundi                      4.71e-08 ***
## country_nameCameroon                      0.00733 ** 
## country_nameCentral African Republic      0.00936 ** 
## country_nameChad                          0.68447    
## country_nameCongo                         0.25430    
## country_nameDjibouti                      0.76455    
## country_nameEthiopia                     3.64e-07 ***
## country_nameGabon                         0.52364    
## country_nameGambia                        0.66560    
## country_nameGhana                         0.45865    
## country_nameGuinea                        0.68803    
## country_nameGuinea-Bissau                 0.96463    
## country_nameIvory Coast                   0.62233    
## country_nameKenya                        5.56e-05 ***
## country_nameLesotho                      5.89e-07 ***
## country_nameLiberia                       0.33617    
## country_nameMadagascar                   3.41e-08 ***
## country_nameMalawi                        0.00165 ** 
## country_nameMali                          0.98613    
## country_nameMauritania                    0.95224    
## country_nameMozambique                    0.00331 ** 
## country_nameNamibia                       0.88980    
## country_nameNiger                         0.93985    
## country_nameNigeria                       0.97536    
## country_nameRwanda                       1.39e-07 ***
## country_nameSenegal                       0.06528 .  
## country_nameSierra Leone                  0.16939    
## country_nameSomalia                       0.83729    
## country_nameSouth Africa                  0.55734    
## country_nameSudan                         0.06356 .  
## country_nameSwaziland                     0.83430    
## country_nameTanzania, United Republic of 1.00e-10 ***
## country_nameTogo                          0.96483    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3709 on 700 degrees of freedom
## Multiple R-Squared: 0.3391,  Adjusted R-squared: 0.2995 
## Wald test: 13.55 on 42 and 700 DF,  p-value: < 2.2e-16
```

Here, we find a an effect at the 0.10 confidence level, but not at the 0.05 level. Notably, including the state fixed effects reduces the p value of LaggedGDPGrowth from 0.26 to 0.06. While this is still not a significant result, it does show that controlling for confounders can significantly improve our certainty of statistical results.

(g)

```
frstage <- ivreg(InternalConflict ~ LaggedRainfallGrowth + InitialGDP + Democracy + Mountains + EthnicFr

rstage <- resid(frstage)

head(tidy(lm(InternalConflict ~ LaggedGDPGrowth + InitialGDP + Democracy + Mountains + EthnicFrac + Reli
```

```
##              term      estimate    std.error  statistic       p.value
## 1     (Intercept)  2.5100367169 0.1042045386  24.087595  8.294953e-94
## 2 LaggedGDPGrowth -0.0448711969 0.0134271274  -3.341831  8.765760e-04
## 3      InitialGDP -1.3878591337 0.0432091287 -32.119582 9.562442e-140
## 4       Democracy  0.0005370009 0.0002275299   2.360133  1.854217e-02
## 5       Mountains  0.0722343450 0.0012296864  58.742087 1.478668e-272
## 6      EthnicFrac -1.0446318870 0.0583997984 -17.887594  3.331869e-59
```

The coefficient in this case is much smaller (-0.045 vs. -2.8), but it is also statistically significant at the alpha = 0.05 level. We handle endogeneity by utilizing the first stage in this second regression, removing the possiblity of correlation with the error term.

Question 2

(a) Bivariate OLS

```
tv <- import("Data/news_study_MAB.dta")

regtv_1 <- tidy(lm(InformationLevel ~ WatchProgram, data = tv))

regtv_1
```

```
##           term  estimate  std.error statistic       p.value
## 1  (Intercept) 3.1567568 0.04270106 73.926887 5.884937e-270
## 2 WatchProgram 0.2963682 0.08422643  3.518708  4.735592e-04
```

These results may be biased by the fact that those who were likely to watch the program already had higher levels of information. Thus, we may be confusing cause for effect in this case.

(b) Controlled Model

```
regtv_2 <- lm(InformationLevel ~ WatchProgram + PoliticalInterest + ReadNews + Education,
                   data = tv)

summary(regtv_2)
```

```
##
## Call:
## lm(formula = InformationLevel ~ WatchProgram + PoliticalInterest +
##     ReadNews + Education, data = tv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5258 -0.5223  0.2404  0.4777  1.9283
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.6942656  0.1640090  10.330  < 2e-16 ***
## WatchProgram      0.2329059  0.0769573   3.026  0.00261 **
## PoliticalInterest 0.2650756  0.0460088   5.761 1.48e-08 ***
## ReadNews          0.1087893  0.0182718   5.954 5.01e-09 ***
## Education         0.0008844  0.0124248   0.071  0.94328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7421 on 488 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.2053, Adjusted R-squared:  0.1987
## F-statistic: 31.51 on 4 and 488 DF,  p-value: < 2.2e-16
```

```
nobs(regtv_2)
```

```
## [1] 493
```

The results are relatively similar, we still have an estimate for WatchProgram which is statistically significant and has a similar substantive effect on InformationLevel. However, we have not defeated endogeneity as we are still confused about the causal direction of many of these variables. Does higher political interest lead to more information or is it the other way around? In this model, we cannot say.

(c)

```
regtv_tst <- lm(WatchProgram ~ TreatmentGroup + PoliticalInterest + ReadNews + Education,
                     data = tv)

tidy(regtv_tst)
```

```
##                term     estimate   std.error  statistic      p.value
## 1       (Intercept) -0.144715179 0.085775678 -1.6871354 9.220479e-02
## 2    TreatmentGroup  0.406446411 0.034296486 11.8509638 1.045858e-28
## 3 PoliticalInterest  0.036886861 0.023493045  1.5701183 1.170240e-01
## 4          ReadNews  0.015588089 0.009254475  1.6843841 9.273530e-02
## 5         Education -0.002065028 0.006345252 -0.3254446 7.449816e-01
```

```
nobs(regtv_tst)
```

```
## [1] 502
```

The assignment variable should be random assigned. It's useful as an instrument, as it introduces a difference in the treatment variable (WatchProgram) without affecting the dependent variable (InformationLevel). Above, I ran a simple OLS between treatment group and the explanatory variable to make sure that it is a strong instrument, which the above test confirms. ###(d)

```
ireg_tv <- ivreg(InformationLevel ~ WatchProgram + PoliticalInterest + ReadNews + Education
      | TreatmentGroup + PoliticalInterest + ReadNews + Education, data = tv)

ireg_tv
```

```
##
## Call:
## ivreg(formula = InformationLevel ~ WatchProgram + PoliticalInterest +     ReadNews + Education | Trea
##
## Coefficients:
##       (Intercept)       WatchProgram  PoliticalInterest
##         1.6887506          0.2913412          0.2640640
##          ReadNews           Education
##         0.1078766          0.0007689
```

```r
summary(ireg_tv)
```

```
##
## Call:
## ivreg(formula = InformationLevel ~ WatchProgram + PoliticalInterest +
##     ReadNews + Education | TreatmentGroup + PoliticalInterest +
##     ReadNews + Education, data = tv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5086 -0.5071  0.2001  0.4929  1.9362
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.6887506  0.1646510  10.257  < 2e-16 ***
## WatchProgram      0.2913412  0.1613956   1.805   0.0717 .
## PoliticalInterest 0.2640640  0.0461014   5.728 1.78e-08 ***
## ReadNews          0.1078766  0.0184163   5.858 8.64e-09 ***
## Education         0.0007689  0.0124353   0.062   0.9507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7426 on 488 degrees of freedom
## Multiple R-Squared: 0.2043,  Adjusted R-squared: 0.1978
## Wald test:    30 on 4 and 488 DF,  p-value: < 2.2e-16
```

```r
nobs(ireg_tv)
```

```
## [1] 493
```

There are 9 more observations in part (c) than in this 2SLS. This is because there are several information levels which are missing. We therefore find a different result for the first stage of the 2SLS.

(e)

The results suggest that there is not a meaningful correlation between watching the prgram and information level, at the alpha equals 0.05 level. We find less significant results than in part (b). While we cannot say for certain whether we have defeated endogeneity, using the assignment of individuals to groups as an IV helps reduce endogeneity in this experiment as we can understand the difference of proportion in the groups from who watched the program and who didn't as the result of the assignment, thus creating a useful instrumental variable.

Question 4

(a)

```r
inmates <- import("Data/inmates.dta")

inmates <- inmates %>%
  mutate(state = factor(state), year = factor(year))

regin_1 <- tidy(lm(prison ~ educ + age + AfAm + state + year, data = inmates))

head(regin_1)
```

```
##          term      estimate    std.error     statistic      p.value
## 1 (Intercept)  2.911017e-02 4.082012e-04   71.31330309 0.000000e+00
## 2        educ -1.198227e-03 1.391285e-05  -86.12375996 0.000000e+00
## 3         age -3.748158e-04 3.654210e-06 -102.57097019 0.000000e+00
## 4        AfAm  2.117762e-02 1.413494e-04  149.82461390 0.000000e+00
## 5      state2 -9.472433e-05 1.069555e-03   -0.08856425 9.294282e-01
## 6      state4  5.170998e-03 5.231901e-04    9.88359148 4.907607e-23
```

We find that education is extremely significant in the likelihood of going to prison. The coefficient is small, however, when considering the unit of educ is in years, we find that the difference between no education and a full 12 years of education is quite substantive in terms of decreasing the probability of going to prison, reducing the relative probability by more than 30 percent.

(b)

No, we cannot causally conclude that increasing education will reduce crime. There are many confounding factors which alter one's chances of entering into the carceral state. For example, the sample regression does not include economic factors, family background, geograpy, etc… which would bias the regression.

(c)

```r
ftest <- lm(educ ~ age + AfAm + state + year + ca9 + ca10 + ca11, data = inmates)

linearHypothesis(ftest, c("ca9", "ca10", "ca11"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## ca9 = 0
## ca10 = 0
## ca11 = 0
##
## Model 1: restricted model
## Model 2: educ ~ age + AfAm + state + year + ca9 + ca10 + ca11
##
##    Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1 3610612 33743214
## 2 3610609 33600250  3    142964 5120.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the F test, we find that these are strong instruments for the given data.

(d)

```r
ivreg(prison ~ educ + age + AfAm + state + year
           | ca9 + ca10 + ca11 + age + AfAm + state + year,
           data = inmates)
```

```
## Error: cannot allocate vector of size 1.6 Gb
```

Unfortunately, it is not possible to currently run this regression due to a computer error which I have been unable to resolve. While I should have the RAM available to run the regression, and I have allocated enough to R in order to do so, the ivreg simply will not work for the given size of the data set. Likewise, this prevents me from answering part (e) of this question. #Question 5 ###(a)

```r
demin <- import("Data/democracy_income.csv") %>%
  group_by(CountryCode) %>%
  select(CountryCode, year, democracy_fh, log_gdp, worldincome, YearOrder)

pdemin <- pdata.frame(demin)

pdemin$lag_gdp <- lag(pdemin$log_gdp, k = 1)

tidy(plm(democracy_fh ~ lag_gdp, data = pdemin, model = "pooling"))
```

```
##          term   estimate   std.error statistic       p.value
## 1 (Intercept) -1.3372658 0.073522000 -18.18865  3.094652e-63
## 2     lag_gdp  0.2337698 0.008984252  26.01995 7.472710e-112
```

We find that the lag of gdp is highly significant in the pooled regression model. However, bias remains a concern. We cannot be sure that gdp growth is the only factor leading to democracy. Other theories may point toward institutions or history which are specific to these countries and may influence both gdp growth and democracy.

(b)

```r
head(tidy(plm(democracy_fh ~ lag_gdp + year + CountryCode, data = pdemin, model = "pooling")))
```

```
##          term     estimate  std.error  statistic      p.value
## 1 (Intercept) -0.193512460 0.21733121 -0.8904035 0.3735263226
## 2     lag_gdp  0.038408436 0.02899967  1.3244437 0.1857471661
## 3    year1965  0.006163702 0.03606057  0.1709264 0.8643263930
## 4    year1970 -0.120493663 0.03620568 -3.3280321 0.0009160006
## 5    year1975 -0.139872667 0.03699592 -3.7807596 0.0001683989
## 6    year1980 -0.089822072 0.03799717 -2.3639150 0.0183293485
```

We still find a statistically significant correlation between lag_gdp and democratization. However, the significance of the results is heavily reduced, though it remains significant at the alpha equals 0.05 level. ###(c)

```r
tidy(plm(log_gdp ~ worldincome, data = pdemin, model = "pooling"))
```

```
##          term   estimate  std.error  statistic      p.value
## 1 (Intercept) 7.85352282 0.05523148 142.192869 0.000000e+00
## 2 worldincome 0.02605943 0.00389400   6.692201 3.634629e-11
```

The instrument must first satisfy the inclusion condition, which we determine above using a simple regression. The instrument satisfies this condition, as world income is correlated with log gdp at the alpha equals 0.05 level. The second condition is the exclusion restriction, which means that the dependent variable must not be correlated to the instrument, except through the independent variable. However, we can only determine this theoretically, there is no statistical test which can be done in order to do so.

```
head(tidy(plm(democracy_fh ~ lag_gdp + year + CountryCode | lag(worldincome, k = 1) + year + CountryCode
```

```
##          term    estimate  std.error  statistic    p.value
## 1 (Intercept)  1.51163356 0.79580909  1.8994927 0.05787170
## 2     lag_gdp -0.21298773 0.11645759 -1.8288866 0.06780174
## 3    year1965  0.05206503 0.04299022  1.2110902 0.22623096
## 4    year1970 -0.03297144 0.05450022 -0.6049781 0.54537111
## 5    year1975 -0.01995469 0.06616449 -0.3015921 0.76304415
## 6    year1980  0.05824546 0.07725812  0.7539073 0.45113444
```

We find that the coefficient has become negative once we instrumentalize for worldincome, thus implying a negative relationship between lag_gdp and democracy. This is the opposite of both the OLS and panel data results. Likewise, the statistical significance of the result disappears, meaning that when we instrument the data we can no longer reject the null hypothesis that there is no relation between gdp and democratization.