PSET 7

Sarah Van Alsten

3/2/2020

Question 1.

a.

$$\begin{aligned}\hat{\beta}_{IV} &= \frac{cov(Z_i, X_i)}{cov(Z_i, X_i)} \\ &= \frac{\sum_i (z_i - \overline{z})(y_i - \overline{y})}{\sum_i (z_i - \overline{z})(x_i - \overline{x})}\end{aligned}$$

Because z is binary, we can use the number of obs in each level of z to estimate z_i - zbar

$$= \frac{(N_0\overline{y_1} + N_1\overline{y_1})/N - (N_0\overline{y_0} + N_1\overline{y_0})/N}{(N_0\overline{x_1} + N_1\overline{x_1})/N - (N_0\overline{x_0} + N_1\overline{x_0})/N}$$

where N1 is the number of observations where z =1, N0 = number of observations where z =0

$$\begin{aligned}&= \frac{(N_0\overline{y_1} + N_1\overline{y_1} - N_0\overline{y_0} - N_1\overline{y_0})/N}{(N_0\overline{x_1} + N_1\overline{x_1} - N_0\overline{x_0} - N_1\overline{x_0})/N} \\ &= \frac{(N_0 + N_1)(\overline{y_1} - \overline{y_0})/N}{(N_0 + N_1)(\overline{x_1} - \overline{x_0})/N} \\ &= \frac{\overline{y_1} - \overline{y_0}}{\overline{x_1} - \overline{x_0}}\end{aligned}$$

b. i. The treatment effect estimator $\hat{\beta}_{IV}$ is consistent when Z is uncorrellated with the error (exogeneity) and is associated with X (relevance)

ii. Per the specification of OLS in matrix form, when we run a regression of D on Z, the estimate of the coefficient for Z is given by $(Z'Z)^{-1}Z'D = \hat{\pi}$

This means that our estimates of D can be derived using the form $\hat{D} = Z\hat{\pi} = Z(Z'Z)^{-1}Z'D$

To simplify, let $P_Z = Z(Z'Z)^{-1}Z'$ so $\hat{D} = P_Z D$

In the next equation, regressing Y on $\hat{X}$, the estimator for $\hat{\beta}_{2SLS}$ just for the $\hat{D}$ coefficient, in the same matrix form, is $(\hat{D}'\hat{D})^{-1}\hat{D}Y$

Substituting in for $\hat{D}$:

$$\hat{\beta}_{2SLS} = ((P_Z D)' P_Z D)^{-1}(P_Z D)'Y$$

$$\hat{\beta}_{2SLS} = (D' P_Z' P_Z D)^{-1} D' P_Z' Y$$

Since P_Z is symmetric:

$$\hat{\beta}_{2SLS} = (D' P_Z P_Z D)^{-1} D' P_Z Y$$

Because P_Z is idempotent:

$$\hat{\beta}_{2SLS} = (D' P_Z D)^{-1} D' P_Z Y$$

Substituting back in for the orignal values of P_Z

$$\hat{\beta}_{2SLS} = (D'Z(Z'Z)^{-1}Z'D)^{-1}D'Z(Z'Z)^{-1}Z'Y$$

$$\hat{\beta}_{2SLS} = (Z'D)^{-1}(D'Z(Z'Z)^{-1})^{-1}D'Z(Z'Z)^{-1}Z'Y$$

$$\hat{\beta}_{2SLS} = (Z'D)^{-1}(Z'Z)(D'Z)^{-1}D'Z(Z'Z)^{-1}Z'Y$$

The Z'Z and D'Z 's cancel each other out, so:

$$\hat{\beta}_{2SLS} = (Z'D)^{-1}Z'Y$$

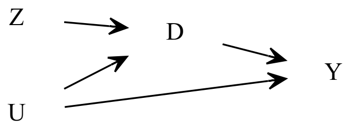
To get all the coefficients for matrix X (not just $\hat{D}$), we follow the same procedure, thus

$$\hat{\beta}_{2SLS} = (Z'X)^{-1}Z'Y$$

Question 2.

```r
#function to compute IV estimator

get_IV <- function(x, z, y){

  #if x,z,y are dataframes, make them into matrices
  if(is.data.frame(x)){
    x <- as.matrix(x)
  }
  if(is.data.frame(y)){
    y <- as.matrix(y)
  }
  if(is.data.frame(z)){
    z <- as.matrix(z)
  }
  ###########################################
  #check if first column is 1's and if not add a column of ones for intercept
  if (all(x[,1] == 1) == FALSE){
    x <- cbind(rep(1, nrow(x)), x)
    warning("Adding col of ones to X matrix")
  }
  if (all(z[,1] == 1) == FALSE){
    z <- cbind(rep(1, nrow(z)), z)
    warning("Adding col of ones to Z matrix")
  }



  #############################################
  #get betas
  #compute IV estimator using OLS formula
  iv.beta <- solve(t(x) %*% z %*% (solve(t(z)%*%z)) %*%  t(z) %*%x) %*%
    (t(x) %*% z %*% (solve(t(z)%*% z)) %*% t(z) %*% y)

  #get error from model
  err <- y - (x %*% iv.beta)
  #get sigma squared
  sigma.squared <- (t(err) %*% err) /(nrow(x) - ncol(x))

  #get var/cov
  var.cov <- sigma.squared[1,1] * solve(t(x) %*% z %*% (solve(t(z) %*% z)) %*%
                                          t(z) %*% x)

  #get std errors
  std.err <- sqrt(diag(var.cov))

  #output betas and ses
  return(list(betas = iv.beta,
              se = std.err))

}
```

Question 3.



a.

It is important to include the unmeasured confounder because without it, there would be no reason to use an instrumental variable. If U did not exist, there would be no open backdoor paths between D and Y, so we could get an unbiased estimate the effect of D on Y by just regressing Y on D. Z only becomes necessary because U confounds the D -> Y path.

b. *Assumption 1:* Z is associated with D (relevance) $Pr(D_i = 1|Z = 1) - Pr(D_i = 1|Z = 0) \neq 0$ This assumption means that early colonist death/illness is associated with the presence/strength of current institutions, and is plausible in AJR, given the authors' results. It also makes theoretical sense- if colonists could not settle somewhere, they would have difficulty establishing institutions which persist into today.

Assumption 2: Z has no effect on Y except through D (exclusion restriction) $Y_i(1, d) = Y_i(0, d)$ for $d = 0, 1$ This assumption means that early colonist death/illness ONLY affects current GDP through the establishment of early institutions. This isn't empirically testable, but the authors do check the robustness of results against a number of possible other ways Z might affect GDP (e.g. environment, colonizer identity) and find that the results remain consistent. This helps allay concerns on the most blatant potential violations of this assumption, so it seems somewhat plausible.

Assumption 3: Z and Y do not share a common cause or are at least independent conditional on other included covariates (independence): $Y^{d,z} \perp\!\!\!\perp Z|X$ for all d, z This assumption means that, given the controlled covariates of latitude, number of European settlers, continent, and latitude, then there are no other unmeasured confounders between colonist deaths and current GDP. This seems somewhat plausible given the examples that local communities are resistant to the diseases that killed early settlers, so shouldn't be affected by the disease or have it affect GDP except through colonization.

Assumption 4: The effect of Z is in the same direction for all individuals/There are no defiers (monotonicity): $D_i(1) \geq D_i(0)$ for all i In this context, this means that increased rates of colonist death always decrease (or at least have 0 effect on) the likelihood of colonization- there are no countries where higher rates of death would make colonists more likely to create institutions. This seems plausible, given that it is unlikely early settlers would waste the time, effort, and lives to try harder amidst higher mortality.

Assumption 5: Z must have a strong association with D: High Cov(Di, Zi) . In this context that means that early settler mortality must have a substantial impact on the development of institutions- if the association is only weak then the estiamtes of the effect of institutions on GDP will be biased. In the context of this study, this is plausible, as the authors demonstrate an F-value of $>$ 10 for the 1st stage IV regression.

c.

```r
#use OLS to estimate effect of avexpr on loggp95
mod1 <- lm(logpgp95 ~ avexpr, data = ajr)
mod1.se <- sqrt(diag(sandwich::vcovHC(mod1, "HC2")))

mod2 <- lm(logpgp95 ~ avexpr + lat_abst + africa + asia + other, data = ajr)
mod2.se <- sqrt(diag(sandwich::vcovHC(mod2, "HC2")))
```

```r
stargazer::stargazer(mod1, mod2, se = list(mod1.se, mod2.se), float = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 09, 2020 - 10:39:20 AM

	Dependent variable:	
	logpgp95	
	(1)	(2)
avexpr	0.522^{***} (0.050)	0.401^{***} (0.066)
lat_abst		0.875 (0.628)
africa		-0.881^{***} (0.154)
asia		-0.577^{*} (0.307)
other		0.107 (0.251)
Constant	4.660^{***} (0.322)	5.737^{***} (0.396)
Observations	64	64
R^2	0.540	0.714
Adjusted R^2	0.533	0.689
Residual Std. Error	0.713 (df = 62)	0.582 (df = 58)
F Statistic	72.816^{***} (df = 1; 62)	28.946^{***} (df = 5; 58)
Note:		*p<0.1; **p<0.05; ***p<0.01

The average protection from expropriation risk is significantly and positively related to current GDP, both before ($\beta = 0.522$, $p < 0.001$) and after ($\beta = 0.401$, $p < 0.001$) adjustment for covariates. This suggests that countries that offer greater protection from expropriation (perhaps, through passage of property rights laws), have better economic outcomes. This is our naive treatment effect estimator. However, we should be concerned about these estimates because there are probably many confounders of expropriation laws and current economic outcome that we have not controlled for. When we omit these confounders, the estimate of the effect of expropriation laws is likely biased. This generally is an identification problem, in that even with an infinite sample size we still have to define the causal structure of associations in order to know that we are truly estimating a causal effect.

d.

```r
#model effect of logem4 on gdp
mod3 <- lm(logpgp95 ~ logem4, data = ajr)
mod3.se <- sqrt(diag(sandwich::vcovHC(mod3, "HC2")))

#control for covariates
mod4 <- lm(logpgp95 ~ logem4+ lat_abst + africa + asia + other, data = ajr)
mod4.se <- sqrt(diag(sandwich::vcovHC(mod4, "HC2")))
```

```r
#make a table
stargazer::stargazer(mod3, mod4, se = list(mod3.se, mod4.se), type = "latex", float = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 09, 2020 - 10:39:20 AM

	Dependent variable:	
	logpgp95	
	(1)	(2)
logem4	-0.573^{***}	-0.377^{***}
	(0.074)	(0.145)
lat_abst		1.046
		(0.886)
africa		-0.723^{***}
		(0.262)
asia		-0.525
		(0.382)
other		0.185
		(0.257)
Constant	10.731^{***}	9.997^{***}
	(0.385)	(0.767)
Observations	64	64
R^2	0.477	0.584
Adjusted R^2	0.469	0.548
Residual Std. Error	0.760 (df = 62)	0.701 (df = 58)
F Statistic	56.603^{***} (df = 1; 62)	16.278^{***} (df = 5; 58)
Note:		*p<0.1; **p<0.05; ***p<0.01

The effect of early settler mortality on current GDP is negative and statistically significant both before (β = -0.573, p $<$ 0.001) and after (β = -0.377, p $<$ 0.001) controlling for covariates. This suggests that countries where more colonizers died have lower earnings in present day than those where fewer died. This estimator purportedly represents the ITT effect, and we can only interpret it as causal if there are no defiers, and if conditional on the included covariates, there is conditional ignorability among those with different instrument values.

```
#use function to estimate LATE
#y = loggp95, z = logem4, x = avexpr

#no covariates
get_IV(z = as.matrix(ajr[, "logem4"]),
       y = as.matrix(ajr[,"logpgp95"]),
       x = as.matrix(ajr[, "avexpr"]))
```

```
## $betas
##         logpgp95
##        1.9096665
## avexpr 0.9442794
##
## $se
##                 avexpr
## 1.0267273 0.1565255
```

```
#with lat_abst, africa, asia, and other.
get_IV(z = as.matrix(ajr[, c("lat_abst", "africa", "asia", "other", "logem4")]),
       y = as.matrix(ajr$logpgp95),
       x = as.matrix(ajr[, c("lat_abst", "africa", "asia", "other", "avexpr")]))
```

```
## $betas
##                 [,1]
##            1.4404521
## lat_abst -1.1781788
## africa   -0.4372669
## asia     -1.0470853
## other    -0.9904017
## avexpr    1.1070772
##
## $se
##            lat_abst    africa      asia     other    avexpr
## 2.8395869 1.7554404 0.4242092 0.5245609 0.9979844 0.4635725
```

```
#Use ivreg function to reestimate effect
library(AER)

#unadjusted
aer.mod <- ivreg(logpgp95 ~ avexpr | logem4,
                 data = ajr)

summary(aer.mod, vcov = sandwich, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = logpgp95 ~ avexpr | logem4, data = ajr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44903 -0.56242  0.07311  0.69564  1.71752
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9097     1.1740   1.627    0.109
## avexpr        0.9443     0.1761   5.362 1.29e-06 ***
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1  62     16.85 0.00012 ***
## Wu-Hausman         1  61     21.82 1.7e-05 ***
## Sargan             0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9483 on 62 degrees of freedom
## Multiple R-Squared: 0.187,   Adjusted R-squared: 0.1739
## Wald test: 28.75 on 1 and 62 DF,  p-value: 1.289e-06
```

```r
#adjusted model
aer.mod2 <- ivreg(logpgp95 ~ avexpr + lat_abst + africa + asia + other|
                    lat_abst + africa + asia + other +logem4 ,
                  data = ajr)

summary(aer.mod2, vcov = sandwich, diagnostics = TRUE)
```

```
## 
## Call:
## ivreg(formula = logpgp95 ~ avexpr + lat_abst + africa + asia + 
##     other | lat_abst + africa + asia + other + logem4, data = ajr)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.7155 -0.6381 -0.1535  0.8188  2.0714 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   1.4405     3.0736   0.469   0.6411  
## avexpr        1.1071     0.5029   2.201   0.0317 *
## lat_abst     -1.1782     1.7917  -0.658   0.5134  
## africa       -0.4373     0.3775  -1.158   0.2515  
## asia         -1.0471     0.5049  -2.074   0.0425 *
## other        -0.9904     1.0594  -0.935   0.3537  
## 
## Diagnostic tests:
##                  df1 df2 statistic p-value  
## Weak instruments   1  58     3.282  0.0752 .
## Wu-Hausman         1  57     5.045  0.0286 *
## Sargan             0  NA        NA      NA  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.082 on 58 degrees of freedom
## Multiple R-Squared: 0.01082, Adjusted R-squared: -0.07445 
## Wald test: 5.689 on 5 and 58 DF,  p-value: 0.000246
```

```r
#make a table
star.out <- stargazer::stargazer(aer.mod, aer.mod2, type = "latex", float = FALSE,
                                 add.lines=list(
                                   c("Model 1 instrument strong",
                                     "F(1,62)=16.85",
                                     "p < 0.001"),
                                   c("Model 2 instrument weak",
                                     "F(1,58) = 3.282",
                                     "p = 0.075")))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 09, 2020 - 10:39:22 AM

	Dependent variable:	
	logpgp95	
	(1)	(2)
avexpr	0.944^{***}	1.107^{**}
	(0.157)	(0.464)
lat_abst		−1.178
		(1.755)
africa		−0.437
		(0.424)
asia		-1.047^{*}
		(0.525)
other		−0.990
		(0.998)
Constant	1.910^{*}	1.440
	(1.027)	(2.840)
Model 1 instrument strong	F(1,62)=16.85	p < 0.001
Model 2 instrument weak	F(1,58) = 3.282	p = 0.075
Observations	64	64
R^2	0.187	0.011
Adjusted R^2	0.174	−0.074
Residual Std. Error	0.948 (df = 62)	1.082 (df = 58)
Note:		*p<0.1; **p<0.05; ***p<0.01

While results show that the instrument of early settler mortality is strong in the unadjusted model, after adjusting for covariates the instrument was only weakly associated with protection against expropriation, suggesting that effect estimates from this model may be biased. Even so, both models estimate that institutions (as instrumented by mortality) do have a positive effect on GDP.