Problem Set 3

MOEC0021 Empirical Methods

Fenqi Guo Wenjie Tu

Fall Semester 2020

Instrumental Variables

1. Theory - Endogenous Instruments

1(a)

z should be exogenous and relevant. These conditions imply $cov(z, \varepsilon) = 0$ and $cov(z, x) \neq 0$.

1(b)

In this context, the exogeneity assumption **cannot** be tested. Therefore the IV strategy in this case must rely on an assumption which cannot be verified with data. Relevance assumption instead can easily be tested in the so-called *first stage*. One only needs to regress the endogenous regressor on the instrument, then check whether the coefficient of z is significantly different from 0.

1(c)

If the square of the *t-statistic* on the z-coefficient is smaller than 10, then z is a weak instrument.

1(d)

The numerator represents the (asymptotic) bias of the IV estimator, while the denominator represents the (asymptotic) bias of the OLS estimator.

1(e)

$$\text{plim}\hat{\beta}_{OLS} = \beta + \frac{cov(x, \varepsilon)}{var(x)} \implies \text{plim}\hat{\beta}_{OLS} - \beta = \frac{cov(x, \varepsilon)}{var(x)}$$

$$\text{plim}\hat{\beta}_{IV} = \beta + \frac{cov(z, \varepsilon)}{cov(z, x)} \implies \text{plim}\hat{\beta}_{IV} - \beta = \frac{cov(z, \varepsilon)}{cov(z, x)}$$

$$\frac{\text{plim}\hat{\beta}_{IV} - \beta}{\text{plim}\hat{\beta}_{OLS} - \beta} = \frac{cov(z, \varepsilon)}{cov(z, x)} \cdot \frac{var(x)}{cov(x, \varepsilon)}$$

1(f)

When the ratio equals 0, it means that the numerator is 0, i.e., the IV estimator is consistent (or asymptotically unbiased). If the ratio is larger than 1 instead, it means that the bias of the IV estimator is larger than that of the OLS estimator.

1(g)

By assumption, $cov(x, \varepsilon) \neq 0$ and $var(x) \neq 0$ (otherwise x would be a constant). For the ratio to be zero, $cov(z, \varepsilon)$ must equal 0 and $cov(z, x) \neq 0$, which are exactly the conditions required for solving the endogeneity problem in x.

1(h)

As discussed on 1(f), the bias in the OLS estimator would be smaller than that of the IV estimator if the ratio is larger than 1. In this case, it would be preferable to discard the IV estimate and just rely on the OLS one, even if biased. Suppose for instance that the instrument is not fully exogenous (i.e., $cov(z, \varepsilon) \neq 0$) and the instrument is weak (i.e., $cov(x, z) \approx 0$). In this case, using z as an instrument for x could lead to an even more severe bias than coming from a least squares estimation.

Empirical Application: IV Regression

2(a)

```
library(stargazer) # print tables
library(AER) # iv regression

# read data
d.twins <- read.table("http://bit.ly/1YATkWe", header = T, sep = ",")
```

```
# generate variables
d.twins$lnearn <- log(d.twins$earning)
d.twins$agesq <- (d.twins$age)^2
```

```
model1.ols <- lm(lnearn ~ highqua + age + agesq, data = d.twins)

## reproduce the results in paper(2003)
# initialize a column with zeros
d.twins$highqua_iv <- numeric(dim(d.twins)[1])

# flip highqua for each twin pair
for (i in 1:dim(d.twins)[1]){
  if (i %% 2 == 1){
    d.twins[i, "highqua_iv"] = d.twins[i+1, "twihigh"]
  }
  else {
    d.twins[i, "highqua_iv"] = d.twins[i-1, "twihigh"]
  }
}
```

```
model1.iv.1 <- ivreg(lnearn ~ highqua + age + agesq |
                       highqua_iv + age + agesq, data = d.twins)

## reproduce the results in paper(2011)
model1.iv.2 <- ivreg(lnearn ~ highqua + age + agesq |
twihigh + age + agesq, data = d.twins)

stargazer(model1.ols, model1.iv.1, model1.iv.2,
          title='Reproduction for Results', header = F, single.row = T,
          model.names = F, keep.stat=c('n', 'rsq'),
          column.labels=c('OLS', 'IV (Paper 2003)', 'IV (Paper 2011)'))
```

Table 1: Reproduction for Results

	Dependent variable:		
	lnearn		
	OLS	IV (Paper 2003)	IV (Paper 2011)
	(1)	(2)	(3)
highqua	0.077*** (0.011)	0.085*** (0.012)	0.087*** (0.017)
age	0.078*** (0.021)	0.077*** (0.021)	0.076*** (0.021)
agesq	−0.001*** (0.0003)	−0.001*** (0.0003)	−0.001*** (0.0003)
Constant	−0.428 (0.435)	−0.531 (0.441)	−0.568 (0.467)
Observations	428	428	428
R^2	0.149	0.148	0.147
Note:			*p<0.1; **p<0.05; ***p<0.01

i There are some discrepancies with the coefficient of *age_squared*. In the original paper the coefficients are -0.097 for OLS and -0.095 for IV, because the author used $\frac{age^2}{100}$, but in our analysis we simply use age^2, so the coefficients are -0.001. We don't think this is serious.

Another discrepancy we regard as relevant. In the column (2), we assume that in the original data set, the variable *twihigh* of twin 1 contains twin 1's estimation of twin 2's education, therefore we generate a new variable *highqua_iv* to reflect this data managing process. This result is the same as the original paper. In the column (3), we assume that the variable *twihigh* of twin 1 contains twin 2's estimation of twin 1's education, so we use this variable directly in the IV regression model. This result is the same as the AER comment paper. We leave this as an open question because the original data set doesn't label this variable clearly. In the following questions (Problem C) we use the first interpretation.

ii There is no important information contained in the constant here. Since the dependent variable is in logs, it has no meaningful interpretation for constant term. The only purpose of a constant in this case is to allow us to rescale the error terms, so they have mean zeros.

iii Interpretation of coefficients

- OLS: one extra year of education is expected to increase individual earnings by 7.7 percent, on average, holding all other factors constant.
- IV: one extra year of education is expected to increase individual earnings by 8.7 percent, on average, holding all other factors constant.

2(b)

i Why years of education (*highqua*) might be endogenous?

- Omitted variable bias: one's innate ability is likely to affect both years of education and income.
- Omitted variable bias: income of one's parents is also likely to affect both years of education and income.
- Measurement error: level of education may be misreported by people with lower levels of education because of embrassement or social stigma.

ii Conjecture about the likely sign of the bias:

- The sign of bias depends on the sign of $\beta_{ability}$ and the sign of $Cov(educ, ability)$. We are more likely to have a positive bias because we expect $\beta_{ability} > 0$ and $Cov(educ, ability) > 0$.
- The sign of bias depends on the sign of $\beta_{parsinc}$ and the sign of $Cov(educ, parsinc)$. We expect a positive bias because both $\beta_{parsinc}$ and $Cov(educ, parsinc)$ are likely to be positive.
- Attenuation bias due measurement error in *educ*. Measurement error pushes the estimated coefficient towards zero, implying that we may have a negative bias.

iii OVB problems: instrumenting a twin's education level with the sibling's estimate of the first twin's education does not resolve the potential endogeneity problem because it does not address the correlation between education and ability.

Measurement error: by asking respondents about their twin's level of education, we are probably trying to control for misreported education levels. In this regard, the variable seems to be both relevant and exogenous. We expect the instrumental variable to be correlated with the endogenous variable. We also expect the instrumental variable to be uncorrelated with the sibling's measurement error and with the sibling's wage, if not through her education level.

iv As claimed, the instrument is not able to address the endogeneity resulting from omitted variables. It can only correct for attenuation bias, if any. From column (1) to column (2) and column (3), we see an increase in the coefficient, which is what we should expect from correcting for attenuation bias.

However, since we believe that there are omitted variables in the estimated model, we expect a positive bias in the OLS estimator of *highqua*. Overall, the IV estimator is even more biased upward than the OLS estimator.

```
## test relevance
model.1ststage <- lm(highqua ~ twihigh + age + agesq, data = d.twins)
model.2ndstage <- lm(lnearn ~ highqua + age + agesq + model.1ststage$residuals,
                     data = d.twins)
stargazer(model.1ststage, model.2ndstage, model1.iv.2, keep.stat = c('n', 'rsq', 'f'),
          title = 'Test relevance', header = F, single.row = T, model.names = F,
          column.labels = c("1st stage", "2nd stage", "IV"))
```

```
# first stage
beta1 <- coef(summary(model.1ststage))[2,1]
beta1.se <- coef(summary(model.1ststage))[2,2]
t.stat <- beta1 / beta1.se
f.stat <- (t.stat)^2

print(paste("The F-statistic is ", round(f.stat, digits=3)))
```

```
## [1] "The F-statistic is  290.168"
```

Table 2: Test relevance

	Dependent variable:		
	highqua	lnearn	
	1st stage	2nd stage	IV
	(1)	(2)	(3)
twihigh	0.631*** (0.037)		
highqua		0.087*** (0.017)	0.087*** (0.017)
age	0.053 (0.076)	0.076*** (0.021)	0.076*** (0.021)
agesq	−0.001 (0.001)	−0.001*** (0.0003)	−0.001*** (0.0003)
residuals		−0.018 (0.022)	
Constant	4.835*** (1.535)	−0.568 (0.467)	−0.568 (0.467)
Observations	428	428	428
R^2	0.446	0.150	0.147
F Statistic	113.802*** (df = 3; 424)	18.699*** (df = 4; 423)	
Note:			*p<0.1; **p<0.05; ***p<0.01

v Given the large F-statistic, we can conclude that the instrument is not weak. (As a rule of thumb, Stock and Watson (1997) define an instrument weak if the *F-statistic* is below 10).

vi We do not really believe these results. Having argued that there might be omitted variables in the model. The instrument is only likely to address the bias caused by measurement error.

2(c)

```
# reshape the data from long to wide
d.twins.wide <- reshape(
  subset(d.twins, select = c("family", "twinno", "earning", "highqua",
                             "twihigh", "age", "lnearn", "agesq")),
  timevar = "twinno", idvar = "family", direction = "wide"
)
```

```
##  [1] "family"    "earning.1" "highqua.1" "twihigh.1" "age.1"     "lnearn.1"
##  [7] "agesq.1"   "earning.2" "highqua.2" "twihigh.2" "age.2"     "lnearn.2"
## [13] "agesq.2"
```

```
# generate variables
d.twins.wide$dlnearn <- d.twins.wide$lnearn.1 - d.twins.wide$lnearn.2
d.twins.wide$dhigh <- d.twins.wide$highqua.1 - d.twins.wide$highqua.2
d.twins.wide$dtwihigh <- d.twins.wide$twihigh.1 - d.twins.wide$twihigh.2
```

```
## regress dlnearn on dhigh without constant
model2.ols <- lm(dlnearn ~ 0 + dhigh, data = d.twins.wide)
model2.iv <- ivreg(dlnearn ~ 0 + dhigh | 0 + dtwihigh, data = d.twins.wide)

stargazer(model2.ols, model2.iv, header = F, keep.stat = c("n", "rsq"),
          title = "Regression results from 2(c)", single.row = T,
          model.names = F, column.labels = c("OLS", "IV"))
```

Table 3: Regression results from 2(c)

	Dependent variable:	
	dlnearn	
	OLS	IV
	(1)	(2)
dhigh	0.039* (0.023)	0.077** (0.033)
Observations	214	214
R^2	0.014	0.001
Note:	*p<0.1; **p<0.05; ***p<0.01	

i Trivially, since they are twins, the difference in age is zero.

```
## redo the estimations with constant
model3.ols <- lm(dlnearn ~ 1 + dhigh, data = d.twins.wide)
model3.iv <- ivreg(dlnearn ~ 1 + dhigh | 1 + dtwihigh, data = d.twins.wide)
stargazer(model2.ols, model2.iv, model3.ols, model3.iv,
          header = F, keep.stat = c("n", "rsq"), model.names = F,
          title = "Regression results from 2(c)", no.space = T,
          column.labels = c("OLS", "IV", "OLS", "IV"))
```

Table 4: Regression results from 2(c)

	Dependent variable:			
	dlnearn			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
dhigh	0.039*	0.077**	0.039*	0.078**
	(0.023)	(0.033)	(0.023)	(0.033)
Constant			0.014	0.013
			(0.047)	(0.047)
Observations	214	214	214	214
R^2	0.014	0.001	0.014	0.0004
Note:			*p<0.1; **p<0.05; ***p<0.01	

ii The constant is excluded in the regression because theoretically, after taking the difference, constants cancel each other out. We run the regressions with constant and see that the constant is not statistically different from zero as expected.

```
stargazer(model1.ols, model2.ols, header = F, keep.stat = c("n", "rsq"),
          model.names = F, column.labels = c("pooled", "within-family"),
```

```
          title = "OLS regressions comparsion between (a) and (c)", no.space = T,
          keep = c("highqua", "dhigh"), single.row = T)
```

Table 5: OLS regressions comparsion between (a) and (c)

	Dependent variable:	
	lnearn pooled	dlnearn within-family
	(1)	(2)
highqua	0.077*** (0.011)	
dhigh		0.039* (0.023)
Observations	428	214
R^2	0.149	0.014
Note:	*p<0.1; **p<0.05; ***p<0.01	

iii In OLS regressions, within estimator is smaller than pooled estimator. The increase in the standard errors can be attributed to the decrease in the number of observations. By taking the difference, we try to mitigate the endogeneity problem caused by the omission of a variable to proxy for ability. Twins should have similar innate ability, so by taking the difference, any correlation between the regressors and the error term should be removed. The OVB for ability should be positive and we therefore expect a corrected coefficient to be smaller.

Notice that this procedure also mitigates the bias coming from omission of family income (and any family-specific omitted variable), for the same reasoning.

```
stargazer(model1.ols, model2.ols, model1.iv.2, model2.iv, header = F,
          keep.stat = c("n", "rsq"), model.names = F, no.space = T,
          column.labels = c("pooled (OLS)", "within (OLS)",
                            "pooled (IV)", "within (IV)"),
          title = "Pooled and within estimates in OLS and IV",
          keep = c("highqua", "dhigh"), single.row = T)
```

Table 6: Pooled and within estimates in OLS and IV

	Dependent variable:			
	lnearn pooled (OLS)	dlnearn within (OLS)	lnearn pooled (IV)	dlnearn within (IV)
	(1)	(2)	(3)	(4)
highqua	0.077*** (0.011)		0.087*** (0.017)	
dhigh		0.039* (0.023)		0.077** (0.033)
Observations	428	214	428	214
R^2	0.149	0.014	0.147	0.001
Note:				*p<0.1; **p<0.05; ***p<0.01

iv In IV estimations, within estimator is smaller than pooled estimator. By instrumenting with the difference in twins' estimated years of schooling, we are trying to remove measurement error along with the

ability (or parents' income) bias. The measurement error should bias the estimate downwards. We therefore expect $\hat{\beta}_{IV} > \hat{\beta}_{OLS}$.

Omitted variable bias and measurement error bias move in opposite directions. Overall the effects seem to cancel out as we can see from column (1) and column (4) though theoretically they do not have to.

v Under this model specification, results are statistically significant and the instrument is strong. More importantly, this model can remove the bias caused by omitted variable such as ability and parents' income (or any family-specific omitted variables). However, we may still have omitted other relevant variables, which may bias our estimates. Such omitted variables could be gender, field of study, sector of occupation (or any individual-specific omitted variables that could affect both earnings and years of schooling).

2(d)

```
# generate variable
d.twins.wide$absearn <- abs(d.twins.wide$earning.1 - d.twins.wide$earning.2)
d.twins.amin <- d.twins.wide[d.twins.wide$absearn <= 60, ]
```

```
model4.ols <- lm(dlnearn ~ 0 + dhigh, data = d.twins.amin)
model4.iv <- ivreg(dlnearn ~ 0 + dhigh | 0 + dtwihigh, data = d.twins.amin)

stargazer(model2.ols, model2.iv, model4.ols, model4.iv, no.space = T,
          keep.stat = c('n', 'rsq'), header = F, model.names = F, single.row = T,
          column.labels = c("OLS (outliers)", "IV (outliers)",
                            "OLS (no outliers)", "IV (no outliers)"))
```

Table 7:

	Dependent variable:			
	dlnearn			
	OLS (outliers)	IV (outliers)	OLS (no outliers)	IV (no outliers)
	(1)	(2)	(3)	(4)
dhigh	0.039* (0.023)	0.077** (0.033)	0.028 (0.019)	0.036 (0.027)
Observations	214	214	210	210
R^2	0.014	0.001	0.011	0.010
Note:				*p<0.1; **p<0.05; ***p<0.01

i The decrease in magnitude and statistical significance of the estimated coefficient is due to the fact that OLS puts more weights on outliers (observations with higher variance). Since the results are not robust to the omission of these outliers, we should cast some doubts on the validity of the instrument variable: the positive association we have found in previous analyses seems to be driven by these 4 outliers. This argument applies irrespective of whether it is reasonable to drop those observations (which depends mostly on whether they are due to errors in reporting).

ii A twin that completed on more year of education than her sibling is expected to earn one average 3.6% more than her sibling's earning.

iii A difference of 60 pounds per hour is significant. This could stem from either misreporting or data entry error but we cannot attribute that to wrong data. As argued before, we should be skeptical about the results found in the paper since they are not robust to the omission of outliers. A well-designed study would show estimates from both regression with outliers and regression without outliers and provide evidence that results are not driven by outliers.