

Computer Exercise 6: Measurement Error & Instrumental Variables

Read in the data set, twins.dta from the course website. The data set contains information on:

hrwage1 the log of hourly wages of twin 1
 male1 = 1 if twin 1 is male, = 0 otherwise
 white1 = 1 if twin 1 is white, = 0 otherwise
 age (agesq) the age (and its square) of twin 1
 educ1 self-reported years of education of twin 1
 educ2 twin 2's estimate of years of education of twin 1

1. Summarise the data, take note of the average values of the variables and their standard errors

```
su educ1 educ2 hrwage1 age male1 white1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
educ1	183	13.96175	2.204805	8	20
educ2	183	13.98361	3.289135	5	26
hrwage1	162	13.77443	16.029	1.666667	149.5193
age	183	38.69402	13.03625	18.78166	79.12389
male1	183	.4590164	.4996846	0	1
white1	183	.9398907	.2383413	0	1

The mean values of the two (mis-measured) education variables are very similar, though, as might be expected, the variation in twin 2's estimate of twin 1's education is larger than the self-reported variable. The mean age of twins is 38.7, 45.9% of the sample are male and 94% are white.

2. Regress the **log** of hourly wages on self-reported years of education of twin1

Interpret your results. In particular what is the estimated effect of years of education?

```
g loghw=log( hrwage1)
(21 missing values generated)
```

```
. reg loghw educ1
```

Source	SS	df	MS	Number of obs = 162		
Model	10.4288934	1	10.4288934	F(1, 160)	=	27.30
Residual	61.1296417	160	.382060261	Prob > F	=	0.0000
				R-squared	=	0.1457
				Adj R-squared	=	0.1404
Total	71.5585351	161	.444462951	Root MSE	=	.61811
loghw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ1	.1170179	.0223975	5.22	0.000	.0727851	.1612508
_cons	.7036513	.3193527	2.20	0.029	.0729611	1.334342

Since this is a semi-log (log-lin) model the coefficients are semi-elasticities
 $b_{educ} = d \log(hw) / d educ$

= % change in hourly wage/100 wrt unit change in education (in this case 1 unit = 1 year since education measured in years) So 1 extra year of education raises the hourly wage by 11.7%

3. Now instrument years of education with the twin sibling's estimate and estimate the model by two stage least squares

```
. ivreg2 loghw (educ1= educ2), first small
```

First-stage regressions

First-stage regression of educ1:

Ordinary Least Squares (OLS) regression

```
-----
Total (centered) SS      = 761.6111111
Total (uncentered) SS  =      32935
Residual SS            = 221.3188515
Number of obs          =      162
F( 1, 160)             = 390.60
Prob > F               = 0.0000
Centered R2            = 0.7094
Uncentered R2          = 0.9933
Root MSE               =      1.2
```

```
-----
      educ1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      educ2 |   .5424319   .0274461    19.76   0.000   .4882286   .5966351
      _cons |   6.428231   .3986596    16.12   0.000   5.640918   7.215545
-----
```

Partial R-squared of excluded instruments: 0.7094

Test of excluded instruments:

```
F( 1, 160) = 390.60
Prob > F    = 0.0000
```

Summary results for first-stage regressions:

Variable	Shea Partial R2	Partial R2	F(1, 160)	P-value
educ1	0.7094	0.7094	390.60	0.0000

The first stage tests indicate that the instrument (educ2) is highly correlated with the suspect variable (educ1) though the coefficient is not equal to one suggesting twin 2 tends to over-estimate the education of twin 1

Instrumental variables (2SLS) regression

```
-----
Total (centered) SS      = 71.55853506
Total (uncentered) SS  = 968.2891427
Residual SS            = 61.16465765
Number of obs          =      162
F( 1, 160)             = 21.66
Prob > F               = 0.0000
Centered R2            = 0.1453
Uncentered R2          = 0.9368
Root MSE               =      .62
```

```
-----
      loghw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      educ1 |   .1237985   .0265996     4.65   0.000   .0712668   .1763302
      _cons |   .6080955   .3779924     1.61   0.110  -.1384024   1.354593
-----
```

Sargan statistic (overidentification test of all instruments): 0.000
(equation exactly identified)

Instrumented: educ1

Instruments: educ2

The second stage estimate suggests that the returns to education are a little higher at around 12.4% for every extra year of education (If there is measurement error then the direction of the coefficient change is as expected. Measurement error leads to attenuation bias).

Since we know (see exercise 6) that given two measures of the same variable that the correlation coefficient is a measure of the reliability ratio (the variance of the true variance to the observed variance) then

```
corr educ1 educ2
(obs=183)
```

	educ1	educ2
educ1	1.0000	
educ2	0.8356	1.0000

and since in the presence of measurement error $p \lim(\hat{\beta}) = b \left[\frac{\sigma_{x'}^2}{\sigma_{x'}^2 + \sigma_w^2} \right] \neq b$

gives an alternative way of estimating the true effect of education = estimated coefficient/.84

$$.117/.84 = .139$$

Now repeat the exercise adding controls for male1, age and agesq. How do the results change?

```
. ivreg2 loghw (educ1= educ2) age agesq male1, first small
```

First-stage regressions

First-stage regression of educ1:

Ordinary Least Squares (OLS) regression

		Number of obs =	162
		F(4, 157) =	98.14
		Prob > F =	0.0000
Total (centered) SS	=	761.6111111	Centered R2 = 0.7143
Total (uncentered) SS	=	32935	Uncentered R2 = 0.9934
Residual SS	=	217.5783607	Root MSE = 1.2

educ1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0083294	.0519755	0.16	0.873	-.094332 .1109909
agesq	-.0002437	.0006229	-0.39	0.696	-.001474 .0009867
male1	.1657098	.1909928	0.87	0.387	-.2115372 .5429568
educ2	.5373103	.0279748	19.21	0.000	.4820548 .5925659
_cons	6.484908	1.08841	5.96	0.000	4.335091 8.634724

Partial R-squared of excluded instruments: 0.7015

Test of excluded instruments:

F(1, 157) = 368.91
 Prob > F = 0.0000

Summary results for first-stage regressions:

Variable	Shea Partial R2	Partial R2	F(1, 157)	P-value
educ1	0.7015	0.7015	368.91	0.0000

In the first stage the correlation with educ1 net of controls is little different to that without controls

Instrumental variables (2SLS) regression

Total (centered) SS	=	71.55853506	Number of obs	=	162
Total (uncentered) SS	=	968.2891427	F(4, 157)	=	16.14
Residual SS	=	49.04066536	Prob > F	=	0.0000
			Centered R2	=	0.3147
			Uncentered R2	=	0.9494
			Root MSE	=	.56

loghw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
educ1	.1095494	.0247179	4.43	0.000	.0607268 .1583721
age	.122278	.0246845	4.95	0.000	.0735214 .1710346
agesq	-.0013193	.000296	-4.46	0.000	-.001904 -.0007345
male1	.261065	.0915311	2.85	0.005	.0802738 .4418562
_cons	-1.869371	.5874774	-3.18	0.002	-3.02975 -.7089921

Sargan statistic (overidentification test of all instruments): 0.000
(equation exactly identified)

Instrumented: educ1
Instruments: educ2 age agesq male1

Compare with OLS estimates including controls

. reg loghw educ1 age agesq male1

Model	22.5197595	4	5.62993987	Number of obs	=	162
Residual	49.0387756	157	.312348889	F(4, 157)	=	18.02
Total	71.5585351	161	.444462951	Prob > F	=	0.0000
				R-squared	=	0.3147
				Adj R-squared	=	0.2972
				Root MSE	=	.55888

loghw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
educ1	.1079392	.0207018	5.21	0.000	.0670492 .1488291
age	.1223539	.0246758	4.96	0.000	.0736145 .1710934
agesq	-.0013205	.0002959	-4.46	0.000	-.0019049 -.0007361
male1	.2624389	.0908011	2.89	0.004	.0830895 .4417883
_cons	-1.848275	.5601896	-3.30	0.001	-2.954756 -.7417949

With controls the effect of education is lower than without, but measurement error effect still works as expected (though in this case makes much less difference to the estimates) when compared with the OLS estimates net of controls