

## Computer Assessed Exercise 1

This exercise is one of the required pieces of work for EC5040.  
It will make up 5% of the total marks for this course.

Read in the data set *comptest0910.dta* from the Ec5040 web site.

This contains data on a cross-section sample of individuals who were asked about their hourly wages and their demographic characteristics.

1. To make your data set (and results) unique, first take a random 90% sample of the data. You can do this by typing the following command in stata

```
sample 90
```

Now summarise the mean values of the variables in the data set. In particular work out the average hourly wage, the mean years of education, the proportion female and the proportion living in london.

(5 marks)

```
sample 90  
(7350 observations deleted)
```

```
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	66148	39.94449	11.47758	16	64
doby	66148	1963.564	11.48375	1939	1988
edage	66148	19.68049	12.59624	14	97
hourpay	66148	10.79218	8.489269	4	577
sex	66148	1.502132	.4999992	1	2
sq1	66148	.1590978	.36577	0	1
sq2	66148	.1563615	.3632005	0	1
sq3	66148	.1728548	.3781246	0	1
sq4	66148	.1694231	.375128	0	1
years	66148	9.905046	17.61771	-99	25
sla15	66148	.328793	.4697781	0	1
sla16	66148	.671207	.4697781	0	1
london	66148	.0893451	.2852434	0	1

So the average hourly wage is £10.79, the mean years of education is 9.9 years, the proportion female is 50.2% and the proportion living in London is 8.9%

2. Examine whether the returns to education have changed over the period.  
Estimate the following equation by OLS

$$\ln(\text{hourpay}) = b_0 + b_1 \text{years} + b_2 \text{age} + b_3 \text{female} + b_4 \text{london} + e$$

- you will have to generate the log hourly wage variable and the female dummy variable

Interpret the meaning of the coefficients on years of education and on female and london

(10 marks)

```
g female=sex==2
g lhw=log(hourpay)

reg lhw yearsed age female london
```

Source	SS	df	MS	Number of obs =	66148
Model	2155.30302	4	538.825754	F( 4, 66143) =	2379.68
Residual	14976.6031	66143	.226427635	Prob > F =	0.0000
				R-squared =	0.1258
				Adj R-squared =	0.1258
				Root MSE =	.47584
Total	17131.9061	66147	.258997476		

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
years	.0050873	.0001089	46.71	0.000	.0048738 .0053007
age	.0048879	.0001674	29.21	0.000	.0045599 .005216
female	-.2156734	.0037042	-58.22	0.000	-.2229336 -.2084132
london	.2761516	.0064952	42.52	0.000	.2634209 .2888822
_cons	2.068388	.0071167	290.64	0.000	2.054439 2.082337

Since this is a semi-log (log-lin) model  $B_i = d \ln Y / dx_i$ , the coefficients on any **continuous** right hand side variable could be interpreted as semi-elasticities so the coefficient measures the percentage change in hourly wages (divided by 100) with respect to a unit change in the right hand side variable

So 1 extra year of education raises hourly pay by 0.5%

The coefficients on any dummy variable approximate this semi-elasticity (since the variable is discrete not continuous)

to calculate effect need to use % change in hourly wage =  $\exp(\hat{\beta}_{female}) - 1$

The estimates suggests that the effect of being female was to reduce the hourly wage by  $\exp(-.216) - 1 = 19.4\%$

The estimates suggests that the effect of living in London was to increase the hourly wage by  $\exp(.276) - 1 = 31.8\%$

3. Some of the data may be coded wrongly or there may be codes for missing values. Make sure you drop these observations before you continue. Give reasons for your decisions.

Now repeat the regression using only valid observations. How does your estimate of the returns to education change?

(10 marks)

There are some observations on years of education that are coded as -99. This is wrong. (Typically "minus" codes are used to indicate missing values). Should drop these.

*N.B. There are some individuals in the data set coded with yearsed=0 as “never had an education” and coded as 97 in the variable edage. However the yearsed variable is correctly coded as zero, so there is no need to drop these observations.*

```
reg lhw yearsed age female london
```

Source	SS	df	MS			
Model	4123.70254	4	1030.92564	Number of obs =	64499	
Residual	12427.6606	64494	.192694833	F( 4, 64494) =	5350.04	
Total	16551.3631	64498	.256618238	Prob > F =	0.0000	
				R-squared =	0.2491	
				Adj R-squared =	0.2491	
				Root MSE =	.43897	

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.0774338	.0006595	117.40	0.000	.0761411	.0787265
age	.009047	.0001591	56.85	0.000	.0087351	.0093589
female	-.2154457	.0034605	-62.26	0.000	-.222282	-.2086631
london	.1743913	.006136	28.42	0.000	.1623648	.1864178
_cons	.986557	.0117905	83.67	0.000	.9634476	1.009666

*running the regression again removing the measurement error in yearsed gives the above. As we would expect with measurement error the variable yearsed was severely attenuated before. The measurement error problem also has implications for the estimates of age and living in London, which are both significantly different from the earlier estimates.*

*The R<sup>2</sup> in this model is also much higher, again reflecting the removal of “noisy” data*

4. Test the hypothesis that the effects of being female and of living in London on pay are zero

Note: full marks will only be given if you demonstrate the result using regression output and not using the stata “test” command, though you can use the latter to check your result

(10 marks)

```
. reg lhw yearsed age
```

Source	SS	df	MS			
Model	3222.70145	2	1611.35072	Number of obs =	64499	
Residual	13328.6617	64496	.206658734	F( 2, 64496) =	7797.16	
Total	16551.3631	64498	.256618238	Prob > F =	0.0000	
				R-squared =	0.1947	
				Adj R-squared =	0.1947	
				Root MSE =	.4546	

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.0809919	.0006743	120.11	0.000	.0796702	.0823136
age	.0094578	.0001646	57.45	0.000	.0091351	.0097805
_cons	.8324788	.0119829	69.47	0.000	.8089922	.8559654

To test the validity of the restriction use the F-test

$$F = \frac{RSS_{restricted} - RSS_{unrestricted} / J}{RSS_{unrestricted} / (N - K_{unrestricted})} \sim F(J, N - K_{unrestricted})$$

$J = 2$  coefficients to test hypothesis of zero (female & London)

$K_{unrestricted} = 5$  (number of unrestricted coefficients including the constant)

$$\text{So } F = \frac{(13328.7 - 12427.7)/2}{12427.7/64494} = \frac{450.5}{.192} \sim F(2, 64499-5)$$

$$= 2337.9 \sim F(2, 64494)$$

95% F critical value at  $F(2, 64494) = 3.0$

So estimated F value >  $F_{critical}$

Hence reject null that the variables London and female have no explanatory power (their true effect is zero)

As a check the stata version of the test is

```
. test female london
```

```
( 1) female = 0  
( 2) london = 0
```

```
F( 2, 64494) = 2337.90  
Prob > F = 0.0000
```

5. Test whether there is evidence of heteroskedasticity in the model. If so adjust the standard errors in the model accordingly.

(10 marks)

Note again that full marks will only be given if you show evidence of the regression based version of the test rather than just using the stata command to run the appropriate test

To test whether the residual variance depends on a set of variables, Z

$$u_i^2 = d_0 + Z_i\delta + e_i \quad (1)$$

Since  $u_i^2$  is unobserved replace with OLS residuals (which are consistent estimates of  $u_i$  if the model is correctly specified)

$$\hat{u}_i^2 = d_0 + Z_i\delta + e_i \quad (2)$$

and estimate (2) by OLS

Can show that

$$N \cdot R^2 \stackrel{asy}{\sim} \chi_r^2$$

where  $r$  = number of right hand side variables in  $Z$  (ie excluding constant)

[this is the **Breusch-Pagan** test for heteroskedasticity ]

If  $N \cdot R^2 > \chi_{critical}^2$  then reject null of homoskedasticity

Stata commands to do this are

```
reg lhw yearsed age female london
predict uhat if e(sample), resid
g uhat2=uhat^2
reg uhat2 yearsed age female london
```

Source	SS	df	MS			
Model	177.627802	4	44.4069505	Number of obs =	64499	
Residual	7949.77474	64494	.123263788	F( 4, 64494) =	360.26	
				Prob > F =	0.0000	
				R-squared =	0.0219	
				Adj R-squared =	0.0218	
Total	8127.40254	64498	.126010148	Root MSE =	.35109	

  

uhat2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.0151436	.0005275	28.71	0.000	.0141097	.0161775
age	.0022811	.0001273	17.92	0.000	.0020316	.0025306
female	-.0440869	.0027677	-15.93	0.000	-.0495116	-.0386622
london	.0440937	.0049076	8.98	0.000	.0344748	.0537125
_cons	-.0737005	.0094301	-7.82	0.000	-.0921834	-.0552175

Hence  $N \cdot R^2 = 64499 \cdot 0.0219 = 1412.5$

Since there are 4 degrees of freedom in this test (the variables yearsed, age, female and London) and the 95% critical value for  $\chi^2_4 = 9.5$

then in this case  $N \cdot R^2 >$  critical value

reject the null of homoskedasticity in the residuals, conclude heteroskedasticity exists

To fix up the standard errors to make them consistent (if not efficient) to heteroskedasticity of unknown form use the “robust” correction

```
. reg lhw yearsed age female london, robust
```

Linear regression				Number of obs =	64499	
				F( 4, 64494) =	4246.23	
				Prob > F =	0.0000	
				R-squared =	0.2491	
				Root MSE =	.43897	

lhw	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.0774338	.0007475	103.59	0.000	.0759687	.0788989
age	.009047	.0001565	57.80	0.000	.0087402	.0093538
female	-.2154457	.0034572	-62.32	0.000	-.2222217	-.2086696
london	.1743913	.0068818	25.34	0.000	.160903	.1878796
_cons	.986557	.0124261	79.39	0.000	.9622018	1.010912

-----

compared with unadjusted standard errors main differences are on the yearsed and female. All variables remain significantly different from zero.

6. Concerns over the possible endogeneity of the years of education variable - caused by omitted variable bias or measurement error – suggest a need to run instrumental variable regressions

You are given two potential instruments based on

a) an exogenous change in the school leaving age in the UK

- you are given a dummy variable indicating whether the individual went to school when the minimum school leaving age was 15 (sla15)

b) the UK school year runs from september to august. Students born after april start school 6 months later than those born earlier in the year and so will receive less compulsory schooling

- you are given a dummy variable denoting whether the individual was born in the 4<sup>th</sup> quarter of the school year (sq4)

do the regression based Wu-Hausman test for endogeneity of years of education. Do this twice using each instrument separately. What do you conclude?

(15 marks)

*The (Durbin-Wu) Hausman test for endogeneity is based on a comparison of the IV and OLS estimates. Under the null hypothesis of no endogeneity both OLS and IV will give consistent estimates of the true coefficient values, but OLS will be the most efficient. If endogeneity is present then only IV is consistent so would expect a big difference in the coefficient estimates*

*The asymptotic equivalent version of the test (Wu-Hausman) is to regress the endogenous variable on the **full set** of exogenous variables ie including the original exogenous right hand side variables, save the residuals and include these as an additional term in the original model. An insignificant t value suggest that the residuals are uncorrelated with the endogenous variables (see lecture notes)*

```
reg yearsed sla15 age female london
```

Source	SS	df	MS			
Model	30458.4003	4	7614.60008	Number of obs =	64499	
Residual	442817.986	64494	6.86603383	F( 4, 64494) =	1109.02	
Total	473276.386	64498	7.33784592	Prob > F =	0.0000	
				R-squared =	0.0644	
				Adj R-squared =	0.0643	
				Root MSE =	2.6203	

  

	yearsed	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sla15		-.1767104	.0366542	-4.82	0.000	-.2485526	-.1048681
age		-.0412742	.0015623	-26.42	0.000	-.0443363	-.0382122

female	-.0775805	.0206541	-3.76	0.000	-.1180625	-.0370985
london	1.471064	.0361671	40.67	0.000	1.400177	1.541952
_cons	14.32707	.0562238	254.82	0.000	14.21687	14.43727

. predict res1, resid

. reg lhw yearsed age female london res1

Source	SS	df	MS	Number of obs = 64499		
Model	4303.03164	5	860.606327	F( 5, 64493) = 4531.48		
Residual	12248.3315	64493	.189917223	Prob > F = 0.0000		
				R-squared = 0.2600		
				Adj R-squared = 0.2599		
				Root MSE = .43579		
Total	16551.3631	64498	.256618238			

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	1.137312	.0344978	32.97	0.000	1.069696	1.204927
age	.0592013	.0016398	36.10	0.000	.0559873	.0624152
female	-.1333863	.0043513	-30.65	0.000	-.1419148	-.1248578
london	-1.386783	.0511691	-27.10	0.000	-1.487075	-1.286492
res1	-1.06026	.034504	-30.73	0.000	-1.127888	-.9926321
_cons	-14.39443	.5006794	-28.75	0.000	-15.37576	-13.4131

*Using the 1<sup>st</sup> potential instrument the Wu-Hausman test indicates that the residual is significant which suggests that there is endogeneity in the years ed variable and so need to instrument*

. reg yearsed sq4 age female london

Source	SS	df	MS	Number of obs = 64499		
Model	31266.4518	4	7816.61294	F( 4, 64494) = 1140.53		
Residual	442009.934	64494	6.85350474	Prob > F = 0.0000		
				R-squared = 0.0661		
				Adj R-squared = 0.0660		
				Root MSE = 2.6179		
Total	473276.386	64498	7.33784592			

  

yearsed	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sq4	-.3263178	.0274626	-11.88	0.000	-.3801446	-.2724911
age	-.0473944	.0009306	-50.93	0.000	-.0492185	-.0455703
female	-.0783659	.0206354	-3.80	0.000	-.1188112	-.0379206
london	1.473487	.0361319	40.78	0.000	1.402669	1.544306
_cons	14.57082	.0413586	352.30	0.000	14.48976	14.65188

. predict res2, resid

. reg lhw yearsed age female london res2

Source	SS	df	MS	Number of obs = 64499		
Model	4123.94917	5	824.789834	F( 5, 64493) = 4280.31		
Residual	12427.414	64493	.192693997	Prob > F = 0.0000		
				R-squared = 0.2492		
				Adj R-squared = 0.2491		
				Root MSE = .43897		
Total	16551.3631	64498	.256618238			

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.0614864	.0141117	4.36	0.000	.0338275	.0891453
age	.0082924	.0006858	12.09	0.000	.0069483	.0096365
female	-.2166804	.0036285	-59.72	0.000	-.2237922	-.2095685
london	.1978814	.0216511	9.14	0.000	.1554452	.2403176
res2	.0159823	.0141271	1.13	0.258	-.0117069	.0436715
_cons	1.217986	.2049049	5.94	0.000	.816372	1.6196

-----

However, using the 2nd potential instrument, the Wu-Hausman test indicates that the residual is **not** significant which suggests that there is **no endogeneity** in the yearsed variable suggesting there is no need to instrument

Moral: Endogeneity test is very sensitive to choice of instrument so be careful

7) run IV regressions using each instrument separately

What do you find? Why?

(Hint use the “ivreg2” command in Stata with the “first” option)

(30 marks)

```
. ivreg2 lhw (yearsied=sla15) age female london, first
```

First-stage regressions

-----

First-stage regression of yearsied:

OLS estimation

-----

Estimates efficient for homoskedasticity only  
 Statistics consistent for homoskedasticity only

Total (centered) SS	=	473276.3861	Number of obs =	64499
Total (uncentered) SS	=	10858870	F( 4, 64494) =	1109.02
Residual SS	=	442817.9858	Prob > F =	0.0000
			Centered R2 =	0.0644
			Uncentered R2 =	0.9592
			Root MSE =	2.62

yearsied	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	-.0412742	.0015623	-26.42	0.000	-.0443363 - .0382122
female	-.0775805	.0206541	-3.76	0.000	-.1180625 - .0370985
london	1.471064	.0361671	40.67	0.000	1.400177 1.541952
sla15	-.1767104	.0366542	-4.82	0.000	-.2485526 - .1048681
_cons	14.32707	.0562238	254.82	0.000	14.21687 14.43727

-----

Included instruments: age female london sla15

-----

Partial R-squared of excluded instruments: 0.0004

Test of excluded instruments:

F( 1, 64494) = 23.24  
 Prob > F = 0.0000

-----

Summary results for first-stage regressions

Variable	Shea Partial R2	Partial R2	F( 1, 64494)	P-value
yearsied	0.0004	0.0004	23.24	0.0000

Underidentification tests

Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)

Ha: matrix has rank=K1 (identified)

Anderson canon. corr. N\*CCEV LM statistic Chi-sq(1)=23.24 P-val=0.0000

Cragg-Donald N\*CDEV Wald statistic Chi-sq(1)=23.24 P-val=0.0000

Weak identification test  
 Ho: equation is weakly identified  
 Cragg-Donald Wald F-statistic 23.24  
 See main output for Cragg-Donald weak id test critical values

Weak-instrument-robust inference  
 Tests of joint significance of endogenous regressors B1 in main equation  
 Ho: B1=0 and overidentifying restrictions are valid  
 Anderson-Rubin Wald test F(1,64494)=894.82 P-val=0.0000  
 Anderson-Rubin Wald test Chi-sq(1)=894.89 P-val=0.0000  
 Stock-Wright LM S statistic Chi-sq(1)=882.64 P-val=0.0000

Number of observations N = 64499  
 Number of regressors K = 5  
 Number of instruments L = 5  
 Number of excluded instruments L1 = 1

IV (2SLS) estimation  
 -----

Estimates efficient for homoskedasticity only  
 Statistics consistent for homoskedasticity only

		Number of obs =	64499
		F( 4, 64494) =	52.92
		Prob > F =	0.0000
		Centered R2 =	-29.8157
		Uncentered R2 =	-0.4954
		Root MSE =	2.812
Total (centered) SS	=	16551.36313	
Total (uncentered) SS	=	341082.5805	
Residual SS	=	510042.651	

lhw	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
years	1.137312	.2226053	5.11	0.000	.7010134	1.57361
age	.0592013	.0105812	5.59	0.000	.0384625	.07994
female	-.1333863	.0280776	-4.75	0.000	-.1884174	-.0783551
london	-1.386783	.3301812	-4.20	0.000	-2.033927	-.7396401
_cons	-14.39443	3.230757	-4.46	0.000	-20.7266	-8.062262

Underidentification test (Anderson canon. corr. LM statistic): 23.236  
 Chi-sq(1) P-val = 0.0000

Weak identification test (Cragg-Donald Wald F statistic): 23.242  
 Stock-Yogo weak ID test critical values: 10% maximal IV size 16.38  
 15% maximal IV size 8.96  
 20% maximal IV size 6.66  
 25% maximal IV size 5.53

Source: Stock-Yogo (2005). Reproduced by permission.

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

Instrumented: yearsed  
 Included instruments: age female london  
 Excluded instruments: sla15

Using the 2<sup>nd</sup> instrument

. ivreg2 lhw (years=eq4) age female london, first

First-stage regressions  
 -----

First-stage regression of yearsed:

OLS estimation  
 -----

Estimates efficient for homoskedasticity only  
 Statistics consistent for homoskedasticity only

		Number of obs =	64499
		F( 4, 64494) =	1140.53
		Prob > F =	0.0000
Total (centered) SS =	473276.3861	Centered R2 =	0.0661
Total (uncentered) SS =	10858870	Uncentered R2 =	0.9593
Residual SS =	442009.9344	Root MSE =	2.618

yearsed	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.0473944	.0009306	-50.93	0.000	-.0492185	-.0455703
female	-.0783659	.0206354	-3.80	0.000	-.1188112	-.0379206
london	1.473487	.0361319	40.78	0.000	1.402669	1.544306
sq4	-.3263178	.0274626	-11.88	0.000	-.3801446	-.2724911
_cons	14.57082	.0413586	352.30	0.000	14.48976	14.65188

Included instruments: age female london sq4

Partial R-squared of excluded instruments: 0.0022

Test of excluded instruments:

F( 1, 64494) = 141.19  
 Prob > F = 0.0000

Summary results for first-stage regressions

Variable	Shea Partial R2	Partial R2	F( 1, 64494)	P-value
yearsed	0.0022	0.0022	141.19	0.0000

Underidentification tests

Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)

Ha: matrix has rank=K1 (identified)

Anderson canon. corr. N\*CCEV LM statistic Chi-sq(1)=140.89 P-val=0.0000

Cragg-Donald N\*CDEV Wald statistic Chi-sq(1)=141.20 P-val=0.0000

Weak identification test

Ho: equation is weakly identified

Cragg-Donald Wald F-statistic 141.19

See main output for Cragg-Donald weak id test critical values

Weak-instrument-robust inference

Tests of joint significance of endogenous regressors B1 in main equation

Ho: B1=0 and overidentifying restrictions are valid

Anderson-Rubin Wald test F(1,64494)=15.65 P-val=0.0001

Anderson-Rubin Wald test Chi-sq(1)=15.65 P-val=0.0001

Stock-Wright LM S statistic Chi-sq(1)=15.64 P-val=0.0001

Number of observations N = 64499

Number of regressors K = 5

Number of instruments L = 5

Number of excluded instruments L1 = 1

IV (2SLS) estimation

Estimates efficient for homoskedasticity only  
 Statistics consistent for homoskedasticity only

		Number of obs =	64499
		F( 4, 64494) =	1891.66
		Prob > F =	0.0000
Total (centered) SS =	16551.36313	Centered R2 =	0.2423
Total (uncentered) SS =	341082.5805	Uncentered R2 =	0.9632
Residual SS =	12540.318	Root MSE =	.4409

lhw	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
years	.0614864	.014175	4.34	0.000	.033704 .0892689
age	.0082924	.0006888	12.04	0.000	.0069423 .0096425
female	-.2166804	.0036448	-59.45	0.000	-.223824 -.2095368
london	.1978814	.0217482	9.10	0.000	.1552556 .2405071
_cons	1.217986	.205824	5.92	0.000	.814578 1.621393

Underidentification test (Anderson canon. corr. LM statistic): 140.891  
Chi-sq(1) P-val = 0.0000

Weak identification test (Cragg-Donald Wald F statistic): 141.188  
Stock-Yogo weak ID test critical values: 10% maximal IV size 16.38  
15% maximal IV size 8.96  
20% maximal IV size 6.66  
25% maximal IV size 5.53

Source: Stock-Yogo (2005). Reproduced by permission.

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

Instrumented: yearsed  
Included instruments: age female london  
Excluded instruments: sq4

*The IV estimates using different instruments are very different*

*In particular the IV estimate using the school leaving age 15 dummy is estimated with a very wide confidence interval – even though it is significantly different from zero – and the point estimate 1.14 looks odd (we normally expect the returns to education to be in the range 0.05-0.15)*

*Even though the sla15 dummy passes the 1<sup>st</sup> stage F test with a value >10 (the “rule of thumb” threshold for a good instrument) it is still not enough to prevent a strange result occurring*

*The reason for the poor performance of this instrument lies in its high correlation with the variable age (over 0.8) The resulting multicollinearity makes it hard to detect an independent effect.*

*Moral: be careful in your choice of instruments. Sometimes the tests and thresholds used to help you are not good enough to help you determine choice of instrument.*

8 Do the asymptotically regression equivalent of the Hausman test of over-identification (only check your answer using the “overid” command).

Would you recommend using all the instruments at the same time? Give reasons for your answer

(10 marks)

Test is a variant of the Hausman test in that can compare  $\hat{\beta}_{IV}^{over}$  based on the full set of instruments with  $\hat{\beta}_{IV}^{just}$  based on a subset of instruments that just identify the

model.  $p \hat{\beta}_{IV}^{just} = \beta$  is a consistent estimator by assumption, but if the additional instruments are invalid then  $p \hat{\lim}(\beta_{IV}^{over}) \neq \beta$

An asymptotic equivalent variation of this test is based on the following.

1) estimate the structural form by IV/2SLS using **all** possible instruments and save the residuals,  $\hat{u}_{IV}$

- 2) Regress  $\hat{u}_{IV}$  on all the exogenous variables in the system  
 3) Under the null that all instruments are uncorrelated with the residuals  $u$  then can show that

$$N \cdot R^2 \sim \chi^2_q$$

Where  $q$  = no. of over-identifying restrictions ( $L_2 - K_2$ )

Again if  $\chi^2 > \chi^2_{critical}$  then reject the null that the over-identifying restrictions (some (but which?) of extra instruments) are valid

```
ivreg lhw (years=sd4 sla15) age female london
```

```
Instrumental variables (2SLS) regression
```

Source	SS	df	MS			
Model	-3553.35758	4	-888.339395	Number of obs = 64499		
Residual	20104.7207	64494	.311730094	F( 4, 64494) = 1216.04		
Total	16551.3631	64498	.256618238	Prob > F = 0.0000		
				R-squared = .		
				Adj R-squared = .		
				Root MSE = .55833		

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years=sd4	.2090795	.0167286	12.50	0.000	.1762915	.2418675
age	.0152766	.0008161	18.72	0.000	.013677	.0168762
female	-.2052532	.0045875	-44.74	0.000	-.2142448	-.1962616
london	-.0195196	.0258176	-0.76	0.450	-.0701222	.0310829
_cons	-.9238898	.2429236	-3.80	0.000	-1.40002	-.4477593

```
Instrumented: years=sd4
Instruments: age female london sq4 sla15
```

```
predict uhat3, resid /* save the residuals from this IV regression */
reg uhat3 age female london sla15 sq4
```

Source	SS	df	MS			
Model	160.489402	5	32.0978803	Number of obs = 64499		
Residual	19944.2313	64493	.30924645	F( 5, 64493) = 103.79		
Total	20104.7207	64498	.311710761	Prob > F = 0.0000		
				R-squared = 0.0080		
				Adj R-squared = 0.0079		
				Root MSE = .5561		

  

uhat3	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0056637	.0003316	17.08	0.000	.0050137	.0063136
female	-1.55e-06	.0043834	-0.00	1.000	-.008593	.0085899
london	-.0018657	.0076756	-0.24	0.808	-.01691	.0131785
sla15	-.1651901	.0077802	-21.23	0.000	-.1804392	-.149941
sq4	.0503077	.0058345	8.62	0.000	.0388721	.0617433
_cons	-.1819681	.0119916	-15.17	0.000	-.2054716	-.1584647

then  $N \cdot R^2 \sim \chi^2_q = 64999 \cdot .008 = 519 \sim \chi^2(1)$

- there is 1 overidentifying restriction (one more instrument than endogenous variable)

*the 95%  $X^2$  critical value for 1 degree of freedom is 3.84*

*Note can check this using the stata commands*

```
. ivreg lhw (years=age4 sla15) age female London  
. overid
```

*Tests of overidentifying restrictions:*

Sargan N*R-sq test	514.874	Chi-sq(1)	P-value = 0.0000
Basman test	518.969	Chi-sq(1)	P-value = 0.0000

*So estimated values exceed critical value, reject null that model that the over-identifying restrictions (extra instruments) are valid in the sense of being uncorrelated with the structural form residuals*

*So conclude that at least one of instruments may not satisfy properties of a good instrument*

*Since the reason for the poor performance of this instrument lies in its high correlation with the variable age (over 0.8) The resulting multicollinearity makes it hard to detect an independent effect. This also has implications for the IV estimate in the over-identified case which is in effect a weighted average of the effect of the 2 instruments.*

*Conclude should probably not use sla15 as an instrument*