

Computer Exercise 7 - Answers

Good practice to summarise the data before you do any regressions. Can see average age of sample is 28. 73% live in urban areas, 44% near a 2 year college. Average years of mother's education is 10.6 years. Average hourly wage is 588 (this is US data so the hourly wage variable is measured in cents and averages \$5.88 an hour)

```
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	1900	2500.921	1491.346	3	5218
nearc2	1900	.44	.4965176	0	1
nearc4	1900	.6878947	.4634745	0	1
educ	1900	13.62895	2.592716	1	18
age	1900	27.92211	3.087326	24	34
fatheduc	1900	10.11368	3.691888	0	18
motheduc	1900	10.61526	3.021056	0	18
ethnic	1900	.1589474	.3657233	0	1
urban	1900	.7284211	.444891	0	1
south	1900	.3784211	.485121	0	1
hwage	1900	588.2095	262.1315	100	2404

```
g lhwage=log(hwage)
```

```
g age2=age^2
```

OLS estimates of the model

$$\ln(\text{wage}) = b_0 + b_1\text{age} + b_2\text{age}^2 + b_3\text{educ} + b_4\text{ethnic} + b_5\text{south} + b_6\text{urban} + u$$

produce the following

```
reg lhwage age age2 educ ethnic south urban
```

Source	SS	df	MS	Number of obs = 1900		
Model	92.1566811	6	15.3594469	F(6, 1893)	=	106.56
Residual	272.842445	1893	.144132301	Prob > F	=	0.0000
				R-squared	=	0.2525
				Adj R-squared	=	0.2501
Total	364.999126	1899	.192205964	Root MSE	=	.37965

lhwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.1413237	.0579648	2.44	0.015	.0276421	.2550053
age2	-.0017252	.0010094	-1.71	0.088	-.0037048	.0002544
educ	.031883	.0034976	9.12	0.000	.0250234	.0387425
ethnic	-.1704538	.0254862	-6.69	0.000	-.2204379	-.1204697
south	-.1118315	.0191644	-5.84	0.000	-.149417	-.0742459
urban	.1712394	.0200696	8.53	0.000	.1318786	.2106003
_cons	3.209143	.8223219	3.90	0.000	1.59639	4.821895

The OLS estimate suggests that an extra year of education is associated with a 3.2% increase in wages, (this is a log-lin regression so the coefficients are

interpreted as semi-elasticities, so $dLwage/dEduc = .032 = \% \text{ change in wage} / 100$ with respect to a unit change in education). t value suggests that this effect is statistically significantly different from zero.

Concerns over the possible endogeneity of the education variable (correlated with missing variables like ability or motivation and therefore correlated with the error term) mean that IV estimation may be more appropriate (assuming a good instrument can be found)

Using residence near a 2 year college (nearc2) when a teenager produces the following

```
. ivreg2 lhwage age age2 (educ=nearc2) ethnic south urban, first
```

First-stage regressions

OLS estimation

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

					Number of obs =	1900
					F(6, 1893) =	26.43
					Prob > F =	0.0000
Total (centered) SS	=	12765.40789			Centered R2 =	0.0773
Total (uncentered) SS	=	365687			Uncentered R2 =	0.9678
Residual SS	=	11778.82297			Root MSE =	2.494

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.070702	.3800597	2.82	0.005	.3253216 1.816082
age2	-.0182119	.0066188	-2.75	0.006	-.0311928 -.0052309
ethnic	-1.296014	.1649703	-7.86	0.000	-1.619556 -.9724711
south	-.3456845	.1265002	-2.73	0.006	-.593779 -.09759
urban	.8532563	.1318629	6.47	0.000	.5946444 1.111868
nearc2	.084487	.1178012	0.72	0.473	-.1465468 .3155207
_cons	-2.216928	5.402961	-0.41	0.682	-12.81331 8.379456

Included instruments: age age2 ethnic south urban nearc2

Partial R-squared of excluded instruments: 0.0003

Test of excluded instruments:

F(1, 1893) = 0.51

Prob > F = 0.4733

Summary results for first-stage regressions

Variable	Shea Partial R2	Partial R2	F(1, 1893)	P-value
educ	0.0003	0.0003	0.51	0.4733

IV (2SLS) estimation

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

					Number of obs =	1900
					F(6, 1893) =	7.08
					Prob > F =	0.0000
Total (centered) SS	=	364.9991261			Centered R2 =	-8.9102
Total (uncentered) SS	=	75383.89199			Uncentered R2 =	0.9520

Total (centered) SS = 364.9991261 Centered R2 = -8.9102
 Total (uncentered) SS = 75383.89199 Uncentered R2 = 0.9520
 Residual SS = 3617.22112 Root MSE = 1.38

lhwage	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
educ	.5646628	.7673363	0.74	0.462	-.9392887	2.068614
age	-.4289501	.8474909	-0.51	0.613	-2.090002	1.232101
age2	.0079722	.0144389	0.55	0.581	-.0203276	.036272
ethnic	.5172396	.999222	0.52	0.605	-1.4412	2.475679
south	.0779296	.2797656	0.28	0.781	-.4704008	.62626
urban	-.2909677	.6699194	-0.43	0.664	-1.603986	1.02205
_cons	4.373862	3.374845	1.30	0.195	-2.240712	10.98844

It isn't. Standard errors almost unaffected by correction for heteroskedasticity (of unknown form)

However since IV estimates are radically different as are the standard errors around the estimates, should check for weakness of instrument. Can do this by looking at 1st stage of 2SLS estimation -the regression of the endogenous variable on the instrument(s).

The 1st stage of the 2SLS regression above indicates that nearc2 and educ are not that closely related. While the t value on nearc2 is greater than two, the F value in the 1st stage regression (0.5) is much less than the rule of thumb cutoff point of 10.

Conclude that nearc2 is a weak instrument for educ and that the IV estimates in this case are no more reliable than the OLS ones.

(you should always check that your instrument is correlated with the endogenous variable using the 1st stage regression)

(Note that the 1st stage F test for a weak instrument is the same as the F test of goodness of fit only in a model where there are no other covariates)

Using a different instrument, mother's education

ivreg2 lhwage age age2 (educ=motheduc) ethnic south urban, first

First-stage regressions

 OLS estimation

Estimates efficient for homoskedasticity only
 Statistics consistent for homoskedasticity only

		Number of obs =	1900
		F(6, 1893) =	88.67
		Prob > F =	0.0000
Total (centered) SS	=	12765.40789	Centered R2 = 0.2194
Total (uncentered) SS	=	365687	Uncentered R2 = 0.9728
Residual SS	=	9964.716433	Root MSE = 2.294

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.8379989	.3497928	2.40	0.017	.151979	1.524019
age2	-.0138884	.0060922	-2.28	0.023	-.0258366	-.0019401

ethnic	-.5597177	.156609	-3.57	0.000	-.866862	-.2525734
south	-.1392769	.1161598	-1.20	0.231	-.3670916	.0885379
urban	.6478669	.1204746	5.38	0.000	.4115899	.8841439
motheduc	.3427952	.0184492	18.58	0.000	.3066123	.3789781
_cons	-2.778572	4.969452	-0.56	0.576	-12.52475	6.967607

Included instruments: age age2 ethnic south urban motheduc

Partial R-squared of excluded instruments: 0.1542

Test of excluded instruments:

F(1, 1893) = 345.23

Prob > F = 0.0000

Summary results for first-stage regressions

Variable	Shea Partial R2	Partial R2	F(1, 1893)	P-value
educ	0.1542	0.1542	345.23	0.0000

IV (2SLS) estimation

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

Total (centered) SS	=	364.9991261	Number of obs	=	1900
Total (uncentered) SS	=	75383.89199	F(6, 1893)	=	96.43
Residual SS	=	275.6387627	Prob > F	=	0.0000
			Centered R2	=	0.2448
			Uncentered R2	=	0.9963
			Root MSE	=	.3809

lh wage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
educ	.0472887	.0089347	5.29	0.000	.0297771 .0648004
age	.1248338	.058815	2.12	0.034	.0095586 .240109
age2	-.0014448	.0010236	-1.41	0.158	-.0034511 .0005615
ethnic	-.1505686	.0276816	-5.44	0.000	-.2048236 -.0963135
south	-.1063444	.0194483	-5.47	0.000	-.1444623 -.0682264
urban	.1578743	.0213596	7.39	0.000	.1160103 .1997384
_cons	3.242821	.8251967	3.93	0.000	1.625466 4.860177

This time the correlation of instrument and endogenous variable as revealed by the 1st stage of the 2sls regression, is much stronger. The t value >21 and the F value=345 which is obviously far higher than the rule of thumb value of 10.

The simple correlation coefficient confirms the strength of the correlation of the endogenous variable with this instrument

corr motheduc educ
(obs=1900)

	motheduc	educ
motheduc	1.0000	
educ	0.4407	1.0000

so relative efficiency $\text{Var}(b_{OLS})/\text{Var}(b_{IV}) = (.4407)^2 = .17$

ie variance of OLS estimator is now just some 5 times smaller than IV using motheduc

As a result, the second stage of the IV is much closer to the original OLS estimate. The IV standard error on educ is still somewhat larger than in OLS,

(which is again what you would expect - see lecture notes). If we take the point IV estimate as true, this suggests that the omitted variable causing endogeneity, (say ability), is negatively correlated with education - (see lecture notes on omitted variable bias). Netting out this effect should raise the (true) coefficient on education.

When searching for more instruments remember that, asymptotically, more instruments means a more efficient IV estimator, but that in small samples more instruments **increases** the bias of the IV estimator. Sample might be large enough to try to follow asymptotic rule, so the next step is to add both instruments.

```
ivreg2 lhwage age age2 (educ=motheduc nearc2) ethnic south urban, first
```

First-stage regressions

OLS estimation

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

		Number of obs =	1900
		F(7, 1892) =	75.98
		Prob > F =	0.0000
Total (centered) SS	=	Centered R2 =	0.2194
Total (uncentered) SS	=	Uncentered R2 =	0.9728
Residual SS	=	Root MSE =	2.295

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.8377041	.3498806	2.39	0.017	.1515117 1.523897
age2	-.0138814	.0060938	-2.28	0.023	-.0258327 -.0019301
ethnic	-.5573671	.1569032	-3.55	0.000	-.8650886 -.2496456
south	-.1426417	.1168933	-1.22	0.223	-.371895 .0866117
urban	.6525095	.1217952	5.36	0.000	.4136426 .8913764
motheduc	.3430673	.0184828	18.56	0.000	.3068184 .3793161
nearc2	-.0284945	.108548	-0.26	0.793	-.2413808 .1843919
_cons	-2.768653	4.970819	-0.56	0.578	-12.51751 6.98021

Included instruments: age age2 ethnic south urban motheduc nearc2

Partial R-squared of excluded instruments: 0.1543

Test of excluded instruments:

F(2, 1892) = 172.57

Prob > F = 0.0000

Summary results for first-stage regressions

Variable	Shea Partial R2	Partial R2	F(2, 1892)	P-value
educ	0.1543	0.1543	172.57	0.0000

IV (2SLS) estimation

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

		Number of obs =	1900
		F(6, 1893) =	96.41
		Prob > F =	0.0000
Total (centered) SS	=	Centered R2 =	0.2451
Total (uncentered) SS	=	Uncentered R2 =	0.9963

Residual SS	=	275.5283366	Root MSE	=	.3808
lh wage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
educ	.0469815	.008932	5.26	0.000	.0294751 .0644879
age	.1251627	.058803	2.13	0.033	.0099109 .2404145
age2	-.0014504	.0010234	-1.42	0.156	-.0034563 .0005555
ethnic	-.1509652	.0276756	-5.45	0.000	-.2052084 -.0967219
south	-.1064538	.0194443	-5.47	0.000	-.144564 -.0683436
urban	.1581409	.021355	7.41	0.000	.1162858 .199996
_cons	3.24215	.8250313	3.93	0.000	1.625118 4.859181

Note IV estimate is little changed compared to that using motheduc as sole instrument. Not surprising since 1st stage of 2sls confirms again that nearc2 is a weak (insignificant) instrument, more so when motheduc is included than before. So shouldn't really be using this instrument.

To do the **Wu-Hausman test** for endogeneity of the education variable

1. regress the potentially endogenous variable on *all* the **exogenous** variables in the *system* (ie those exogenous variables both in and outside the original equation)

```
reg educ age age2 ethnic south urban motheduc nearc2
```

Source	SS	df	MS	Number of obs =	1900
Model	2801.05438	7	400.150625	F(7, 1892) =	75.98
Residual	9964.35352	1892	5.26657163	Prob > F =	0.0000
Total	12765.4079	1899	6.72217372	R-squared =	0.2194
				Adj R-squared =	0.2165
				Root MSE =	2.2949

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.8377041	.3498806	2.39	0.017	.1515117 1.523897
age2	-.0138814	.0060938	-2.28	0.023	-.0258327 -.0019301
ethnic	-.5573671	.1569032	-3.55	0.000	-.8650886 -.2496456
south	-.1426417	.1168933	-1.22	0.223	-.371895 .0866117
urban	.6525095	.1217952	5.36	0.000	.4136426 .8913764
motheduc	.3430673	.0184828	18.56	0.000	.3068184 .3793161
nearc2	-.0284945	.108548	-0.26	0.793	-.2413808 .1843919
_cons	-2.768652	4.970819	-0.56	0.578	-12.51751 6.98021

Save the residuals (or the predicted values) from this regression and include them as an extra variable in the original OLS specification (if there were more than one endogenous right hand side variable, just repeat the above and include as many predicted residual values as there are endogenous variables)

```
predict res, resid
```

```
reg lhw educ age age2 ethnic south urban res
```

Source	SS	df	MS	Number of obs =	1900
Model	92.6466341	7	13.2352334	F(7, 1892) =	91.94
				Prob > F =	0.0000

Residual	272.352492	1892	.14394952		R-squared	=	0.2538
					Adj R-squared	=	0.2511
Total	364.999126	1899	.192205964		Root MSE	=	.37941

lhwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0469815	.0088991	5.28	0.000	.0295283	.0644346
age	.1251627	.0585866	2.14	0.033	.0102615	.2400638
age2	-.0014504	.0010197	-1.42	0.155	-.0034501	.0005494
ethnic	-.1509652	.0275738	-5.47	0.000	-.2050434	-.096887
south	-.1064538	.0193728	-5.50	0.000	-.144448	-.0684596
urban	.1581409	.0212764	7.43	0.000	.1164131	.1998686
res	-.0178527	.0096768	-1.84	0.065	-.0368311	.0011256
_cons	3.24215	.8219951	3.94	0.000	1.630038	4.854262

The t value on reshat (or phat) is above the critical value of statistical significance at the 5% level. (Note that the t values from using either the residuals or the predicted values are the same).

conclude that the null hypothesis of no endogeneity **can not be rejected** at the 5% level.

As a check Stata will do the alternative form of the Hausman test -comparing the difference in OLS and IV coefficients net of sampling variation -If no endogeneity IV and OLS will be consistent but only OLS efficient. If endogeneity present only IV is consistent and the IV and OLS coefficients will be a quite different.

```
quietly ivreg lhwage age age2 (educ=motheduc nearc2) ethnic south urban
```

```
. hausman, save
```

```
. quietly reg lhwage age age2 educ ethnic south urban
```

```
. hausman
```

	---- Coefficients ----		(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
	(b) Consistent	(B) Efficient		
educ	.0469815	.031883	.0150985	.0082367
age	.1251627	.1413237	-.0161611	.0105197
age2	-.0014504	-.0017252	.0002748	.0001802
ethnic	-.1509652	-.1704538	.0194886	.010919
south	-.1064538	-.1118315	.0053777	.0034938
urban	.1581409	.1712394	-.0130985	.0074118

b = consistent under Ho and Ha; obtained from ivreg
B = inconsistent under Ha, efficient under Ho; obtained from regress

Test: Ho: difference in coefficients not systematic

$$\begin{aligned} \text{chi2}(6) &= (b-B)'[(V_b-V_B)^{-1}](b-B) \\ &= 3.36 \\ \text{Prob}>\text{chi2} &= 0.7625 \end{aligned}$$

Again since the estimated chi-squared value does not exceed the critical value at 1 degree of freedom (the number of potentially endogenous variables), the result is to accept the null of no endogeneity.

Can't test exogeneity of all instruments, but if the equation is over-identified (more instruments than endogenous RHS variables can ask whether the additional instruments are exogenous).

first save the residuals from the original IV regression

```
quietly ivreg lhwage age age2 (educ=motheduc nearc2) ethnic south urban
predict resiv, resid
```

Then regress these residuals on **all** the **exogenous** variables in the **system** (inside and outside the original equation)

```
reg resiv age age2 ethnic south urban motheduc nearc2
```

Source	SS	df	MS	Number of obs = 1900		
Model	.859253868	7	.122750553	F(7, 1892)	=	0.85
Residual	274.669082	1892	.145173934	Prob > F	=	0.5494
-----				R-squared	=	0.0031
Total	275.528336	1899	.145091278	Adj R-squared	=	-0.0006
-----				Root MSE	=	.38102

resiv	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0003822	.0580898	0.01	0.995	-.1135447	.1143091
age2	-9.36e-06	.0010117	-0.01	0.993	-.0019936	.0019749
ethnic	-.0033919	.0260503	-0.13	0.896	-.0544821	.0476984
south	.0052436	.0194075	0.27	0.787	-.0328188	.043306
urban	-.0072104	.0202214	-0.36	0.721	-.0468689	.0324481
motheduc	-.0003133	.0030687	-0.10	0.919	-.0063316	.005705
nearc2	.0438405	.018022	2.43	0.015	.0084955	.0791855
_cons	-.0154455	.8252932	-0.02	0.985	-1.634026	1.603135

The test is an LM test of the form

$$N \cdot R^2_{\text{auxillary}} = 1900 \cdot 0.0031 = 5.9 \sim \chi^2(L-k)$$

Where the degrees of freedom equals the total number of exogenous variables in the system (9) minus the number of rhs variables in the original regression excluding the constant (6) = number of extra instruments at your disposal

So in this case we have used 2 instruments when we could have proceeded with just 1, so the degrees of freedom for the test are 1

From tables $\chi^2_{(1)} \text{critical} = 3.84$ so reject null that additional instruments are uncorrelated with the structural form residuals.

Note that this result is driven by the significance of nearc2 in the auxillary regression. So may be better to drop this variable.

Hausman version of this test given by

```
quietly ivreg lhwage age age2 (educ=motheduc) ethnic south urban
```

```
. hausman, save
```

You used the old syntax of hausman. [Click here to learn about the new syntax.](#)

(storing estimation results as `_HAUSMAN`)

```
. quietly ivreg lhwage age age2 (educ=motheduc nearc2) ethnic south urban
```

```
. hausman
```

You used the old syntax of hausman. [Click here to learn about the new syntax.](#)

	---- Coefficients ----		(b-B)	sqrt(diag(V_b-V_B))
	(b)	(B)	Difference	S.E.
	Consistent	Efficient		
educ	.0472887	.0469815	.0003073	.0002193
age	.1248338	.1251627	-.0003289	.0011871
age2	-.0014448	-.0014504	5.59e-06	.0000207
ethnic	-.1505686	-.1509652	.0003966	.0005786
south	-.1063444	-.1064538	.0001094	.0003926
urban	.1578743	.1581409	-.0002666	.0004421

b = consistent under Ho and Ha; obtained from ivreg
B = inconsistent under Ha, efficient under Ho; obtained from ivreg

Test: Ho: difference in coefficients not systematic

$$\begin{aligned} \text{chi2}(6) &= (b-B)'[(V_b-V_B)^{-1}](b-B) \\ &= 1.96 \\ \text{Prob}>\text{chi2} &= 0.9231 \end{aligned}$$

Note test can vary (in finite samples) depending on order in which compare variables

```
quietly ivreg lhwage age age2 (educ=nearc2) ethnic south urban
```

```
. hausman, save
```

You used the old syntax of hausman. [Click here to learn about the new syntax.](#)

```
. quietly ivreg lhwage age age2 (educ=motheduc nearc2) ethnic south urban
```

```
. hausman
```

You used the old syntax of hausman. [Click here to learn about the new syntax.](#)

	---- Coefficients ----		(b-B)	sqrt(diag(V_b-V_B))
	(b)	(B)	Difference	S.E.
	Consistent	Efficient		
educ	.5646625	.0469815	.517681	.772622
age	-.4289498	.1251627	-.5541125	.8514105
age2	.0079722	-.0014504	.0094225	.014498
ethnic	.5172392	-.1509652	.6682044	1.001129
south	.0779295	-.1064538	.1843833	.2832077
urban	-.2909674	.1581409	-.4491083	.6738657

b = consistent under Ho and Ha; obtained from ivreg
B = inconsistent under Ha, efficient under Ho; obtained from ivreg

Test: Ho: difference in coefficients not systematic

$$\begin{aligned} \text{chi2}(6) &= (b-B)'[(V_b-V_B)^{-1}](b-B) \\ &= 0.45 \\ \text{Prob}>\text{chi2} &= 0.9984 \end{aligned}$$