Utrecht University

Tutorials Week 7

Pdf file on Blackboard	Dataset on Blackboard	Papers	Description
C.15.3	card.dta	Card, D.(1993): Using geographic variation in college proximity to estimate the return to schooling. No 4483, NBER Working Papers, National Bureau of Economic Research, Inc.	Omitted Variable Bias
C.15.5	card.dta	Card, D.(1993): Using geographic variation in college proximity to estimate the return to schooling. No 4483, NBER Working Papers, National Bureau of Economic Research, Inc.	Difference between OLS and 2SLS, application of the Sargan test to verify for the overidentification restriction.
C.16.2	mroz.dta	T.A. Mroz (1987), “The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions,” Econometrica 55, 765-799	2SLS, test for overidentifying restriction. Simultaneous Equations

C3 Use the data in CARD.RAW for this exercise.

(i) The equation we estimated in Example 15.4 can be written as

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \ldots + u,$$

where the other explanatory variables are listed in Table 15.1. In order for IV to be consistent, the IV for *educ*, *nearc4*, must be uncorrelated with u. Could *nearc4* be correlated with things in the error term, such as unobserved ability? Explain.

(ii) For a subsample of the men in the data set, an IQ score is available. Regress *IQ* on *nearc4* to check whether average IQ scores vary by whether the man grew up near a four-year college. What do you conclude?

(iii) Now, regress *IQ* on *nearc4*, *smsa66*, and the 1966 regional dummy variables *reg662*, …, *reg669*. Are *IQ* and *nearc4* related after the geographic dummy variables have been partialled out? Reconcile this with your findings from part (ii).

(iv) From parts (ii) and (iii), what do you conclude about the importance of controlling for *smsa66* and the 1966 regional dummies in the log(*wage*) equation?

EXAMPLE 15.4 USING COLLEGE PROXIMITY AS AN IV FOR EDUCATION

Card (1995) used wage and education data for a sample of men in 1976 to estimate the return to education. He used a dummy variable for whether someone grew up near a four-year college (*nearc4*) as an instrumental variable for education. In a log(*wage*) equation, he included other standard controls: experience, a black dummy variable, dummy variables for living in an SMSA and living in the South, and a full set of regional dummy variables and an SMSA dummy for where the man was living in 1966. In order for *nearc4* to be a valid instrument, it must be uncorrelated with the error term in the wage equation—we assume this—and it must be partially correlated with *educ*. To check the latter requirement, we regress *educ* on *nearc4* and all of the exogenous variables appearing in the equation. (That is, we estimate the reduced form for *educ*.) Using the data in CARD.RAW, we obtain, in condensed form,

$$\underset{(.24)}{educ} = 16.64 + \underset{(.088)}{.320}\, nearc4 - \underset{(.034)}{.413}\, exper + \ldots$$

$$n = 3{,}010,\ R^2 = .477.$$

We are interested in the coefficient and t statistic on *nearc4*. The coefficient implies that in 1976, other things being fixed (experience, race, region, and so on), people who lived near a college in 1966 had, on average, about one-third of a year more education than those who did not grow up near a college. The t statistic on *nearc4* is 3.64, which gives a p-value that is zero in the first three decimals. Therefore, if *nearc4* is uncorrelated with unobserved factors in the error term, we can use *nearc4* as an IV for *educ*.

The OLS and IV estimates are given in Table 15.1. Interestingly, the IV estimate of the return to education is almost twice as large as the OLS estimate, but the standard error of the IV estimate is over 18 times larger than the OLS standard error. The 95% confidence interval for the IV estimate is between .024 and .239, which is a very wide range. The presence of larger confidence intervals is a price we must pay to get a consistent estimator of the return to education when we think *educ* is endogenous.

nearc4: =1 if near to a 4-year college, 1966
smsa: =1 if in SMSA, 1976
educ=years of schooling

SMSA: Standard Metropolitan Statistical Areas; equivalent to Metropolitan Statistical Areas (MSA)

TABLE 15.1 Dependent Variable: log(*wage*)

Explanatory Variables	OLS	IV
educ	.075 (.003)	.132 (.055)
exper	.085 (.007)	.108 (.024)
$exper^2$	−.0023 (.0003)	−.0023 (.0003)
black	−.199 (.018)	−.147 (.054)
smsa	.136 (.020)	.112 (.032)
south	−.148 (.026)	−.145 (.027)
Observations *R*-squared	3,010 .300	3,010 .238
Other controls: *smsa66, reg662, …, reg669*		

As discussed earlier, we should not make anything of the smaller *R*-squared in the IV estimation: by definition, the OLS *R*-squared will always be larger because OLS minimizes the sum of squared residuals.

i) IQ scores might vary by geographic region, and so does the availability of four-year colleges.
It could be that, for various reasons, people with higher abilities grow up in areas with four-year colleges nearby.

ii) The regression of *IQ* on *nearc4* gives

```
. reg IQ nearc4

      Source |       SS           df       MS      Number of obs   =     2,061
-------------+----------------------------------   F(1, 2059)      =     12.13
       Model |  2869.62905         1  2869.62905   Prob > F        =    0.0005
    Residual |  487188.423     2,059  236.614096   R-squared       =    0.0059
-------------+----------------------------------   Adj R-squared   =    0.0054
       Total |  490058.052     2,060  237.892258   Root MSE        =    15.382

------------------------------------------------------------------------------
          IQ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      nearc4 |     2.5962   .7454966     3.48   0.001     1.134195    4.058206
       _cons |   100.6106   .6274557   160.35   0.000     99.38014    101.8412
------------------------------------------------------------------------------
```

- This shows that the predicted *IQ* score is about 2.6 points higher for a man who grew up near a four-year college. The difference is statistically significant (*t* statistic=3.48; P-value 0.001).
- The average IQ scores vary by whether the man grew up near a four-year college.

(iii) When we add *smsa66*, *reg662*, , *reg669* to the regression in part (ii), we obtain

```
. reg IQ nearc4 smsa66 reg662-reg669
```

Source	SS	df	MS			
Model	30699.1017	10	3069.91017	Number of obs	=	2,061
Residual	459358.951	2,050	224.077537	F(10, 2050)	=	13.70
				Prob > F	=	0.0000
				R-squared	=	0.0626
				Adj R-squared	=	0.0581
Total	490058.052	2,060	237.892258	Root MSE	=	14.969

IQ	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
nearc4	.3478974	.8144087	0.43	0.669	-1.249257	1.945052
smsa66	1.089165	.8086998	1.35	0.178	-.4967934	2.675124
reg662	1.099282	1.649748	0.67	0.505	-2.136074	4.334639
reg663	-1.559295	1.622997	-0.96	0.337	-4.742191	1.6236
reg664	-.5425011	1.916258	-0.28	0.777	-4.300517	3.215515
reg665	-8.47546	1.665513	-5.09	0.000	-11.74173	-5.209185
reg666	-7.421172	1.973869	-3.76	0.000	-11.29217	-3.550175
reg667	-8.39441	1.829768	-4.59	0.000	-11.98281	-4.806013
reg668	-2.924975	2.34463	-1.25	0.212	-7.52308	1.67313
reg669	-2.891917	1.797382	-1.61	0.108	-6.416801	.6329674
_cons	104.7735	1.624972	64.48	0.000	101.5867	107.9602

where, for brevity, the coefficients on the regional dummies are not reported. Now, the relationship between *IQ* and *nearc4* is much weaker and statistically insignificant.
In other words, once we control for region and environment while growing up, there is no apparent link between IQ score and living near a four-year college.

(iv) The findings from parts (ii) and (iii) show that it is important to include *smsa66*, *reg662*,…, *reg669* in the wage equation to control for differences in access to colleges that might also be correlated with ability.

The importance of controlling for smsa66 and the 1966 regional dummies in the log(wage) equation helped to mitigate omitted bias, accounts for regional differences, and provides a better understanding of how education and geographic factors interact to influence IQ and wages.

Utrecht University

C5 Use the data in CARD.RAW for this exercise.

(i) In Table 15.1, the difference between the IV and OLS estimates of the return to education is economically important. Obtain the reduced form residuals, $\hat{v}_2$, from the reduced form regression *educ* on *nearc*4, *exper*, *exper*^{2}, *black*, *smsa*, *south*, *smsa*66, *reg*662, ..., *reg*669—see Table15.1. Use these to test whether *educ* is exogenous; that is, determine if the difference between OLS and IV is *statistically* significant.

(ii) Estimate the equation by 2SLS, adding *nearc2* as an instrument. Does the coefficient on *educ* change much?

(iii) Test the single overidentifying restriction from part (ii).

Structural Equation: $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 expersq + \beta_4 black + \beta_5 smsa + \beta_6 ssouth + \beta_7 smsa66_i + \beta_8 reg_{662} + \cdots \beta_{15} reg_{669+} u_i$

Reduced Form Equation: $educ_i = \alpha_0 + \alpha_1 nearc4 + \alpha_2 exper + \alpha_3 expersq + \alpha_4 black + \alpha_5 smsa + \alpha_5 south + \cdots + v_1$

OLS

. reg lwage educ exper expersq black smsa south smsa66 reg662-reg669

Source	SS	df	MS
Model	177.695591	15	11.8463727
Residual	414.946054	2,994	.138592536
Total	592.641645	3,009	.196956346

Number of obs	=	3,010
F(15, 2994)	=	85.48
Prob > F	=	0.0000
R-squared	=	0.2998
Adj R-squared	=	0.2963
Root MSE	=	.37228

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0746933	.0034983	21.35	0.000	.0678339	.0815527
exper	.084832	.0066242	12.81	0.000	.0718435	.0978205
expersq	-.002287	.0003166	-7.22	0.000	-.0029079	-.0016662
black	-.1990123	.0182483	-10.91	0.000	-.2347927	-.1632318
smsa	.1363845	.0201005	6.79	0.000	.0969724	.1757967
south	-.147955	.0259799	-5.69	0.000	-.1988952	-.0970148
smsa66	.0262417	.0194477	1.35	0.177	-.0118905	.0643739
reg662	.0963672	.0358979	2.68	0.007	.0259801	.1667542
reg663	.14454	.0351244	4.12	0.000	.0756696	.2134105
reg664	.0550756	.0416573	1.32	0.186	-.0266043	.1367554
reg665	.1280248	.0418395	3.06	0.002	.0459878	.2100618
reg666	.1405174	.0452469	3.11	0.002	.0517992	.2292356
reg667	.117981	.0448025	2.63	0.008	.0301343	.2058277
reg668	-.0564361	.0512579	-1.10	0.271	-.1569404	.0440682
reg669	.1185698	.0388301	3.05	0.002	.0424335	.194706
_cons	4.620807	.0742327	62.25	0.000	4.475254	4.766359

IV

Instrumental variables 2SLS regression

Number of obs	=	3,010
Wald chi2(15)	=	769.20
Prob > chi2	=	0.0000
R-squared	=	0.2382
Root MSE	=	.3873

lwage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
educ	.1315038	.0548174	2.40	0.016	.0240637	.238944
exper	.1082711	.0235956	4.59	0.000	.0620246	.1545176
expersq	-.0023349	.0003326	-7.02	0.000	-.0029868	-.001683
black	-.1467757	.0537564	-2.73	0.006	-.2521364	-.0414151
smsa	.1118083	.0315777	3.54	0.000	.0499171	.1736995
south	-.1446715	.027212	-5.32	0.000	-.1980061	-.0913369
smsa66	.0185311	.0215511	0.86	0.390	-.0237082	.0607704
reg662	.1007678	.0375854	2.68	0.007	.0271017	.1744339
reg663	.1482588	.0367162	4.04	0.000	.0762964	.2202211
reg664	.0498971	.0436234	1.14	0.253	-.0356032	.1353974
reg665	.1462719	.0469387	3.12	0.002	.0542738	.2382701
reg666	.1629029	.0517714	3.15	0.002	.0614328	.2643731
reg667	.1345722	.0492708	2.73	0.006	.0380032	.2311413
reg668	-.083077	.0591735	-1.40	0.160	-.1990548	.0329008
reg669	.1078142	.0417024	2.59	0.010	.026079	.1895494
_cons	3.666151	.9223682	3.97	0.000	1.858342	5.473959

Endogenous: educ
Exogenous: exper expersq black smsa south smsa66 reg662 reg663 reg664 reg665 reg666 reg667 reg668 reg669 nearc4

(i) We have to obtain the $\widehat{v_2}$ from the reduced form regression *educ* on *nearc*4, *exper*, *exper*2, *black*, *smsa*, *south*, *smsa*66, *reg*662, ..., *reg*669 and test whether educ is exogenous:

```
. reg educ nearc4 exper expersq black smsa south smsa66 reg662-reg669
```

Source	SS	df	MS		
Model	10287.6179	15	685.841194	Number of obs =	3,010
Residual	11274.4622	2,994	3.76568542	F(15, 2994) =	182.13
Total	21562.0801	3,009	7.16586243	Prob > F =	0.0000
				R-squared =	0.4771
				Adj R-squared =	0.4745
				Root MSE =	1.9405

educ	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
nearc4	.3198989	.0878638	3.64	0.000	.1476194	.4921785
exper	-.4125334	.0336996	-12.24	0.000	-.4786101	-.3464566
expersq	.0008686	.0016504	0.53	0.599	-.0023674	.0041046
black	-.9355287	.0937348	-9.98	0.000	-1.11932	-.7517377
smsa	.4021825	.1048112	3.84	0.000	.1966732	.6076918
south	-.0516126	.1354284	-0.38	0.703	-.3171548	.2139296
smsa66	.0254805	.1057692	0.24	0.810	-.1819071	.2328682
reg662	-.0786363	.1871154	-0.42	0.674	-.4455241	.2882514
reg663	-.027939	.1833745	-0.15	0.879	-.3874918	.3316139
reg664	.117182	.2172531	0.54	0.590	-.3087984	.5431624
reg665	-.2726165	.2184204	-1.25	0.212	-.7008858	.1556528
reg666	-.3028147	.2370712	-1.28	0.202	-.7676536	.1620242
reg667	-.2168177	.2343879	-0.93	0.355	-.6763953	.2427598
reg668	.5238914	.2674749	1.96	0.050	-.0005618	1.048344
reg669	.210271	.2024568	1.04	0.299	-.1866975	.6072395
_cons	16.63825	.2406297	69.14	0.000	16.16644	17.11007

```
. predict v2, res
```

```
. reg lwage educ exper expersq black smsa south smsa66 reg662-reg669 v2
```

Source	SS	df	MS		
Model	177.857408	16	11.116088	Number of obs =	3,010
Residual	414.784236	2,993	.138584777	F(16, 2993) =	80.21
Total	592.641645	3,009	.196956346	Prob > F =	0.0000
				R-squared =	0.3001
				Adj R-squared =	0.2964
				Root MSE =	.37227

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.1315038	.0526906	2.50	0.013	.0281904	.2348172
exper	.1082711	.0226801	4.77	0.000	.0638008	.1527413
expersq	-.0023349	.0003197	-7.30	0.000	-.0029618	-.0017081
black	-.1467758	.0516708	-2.84	0.005	-.2480896	-.045462
smsa	.1118083	.0303526	3.68	0.000	.0522943	.1713223
south	-.1446715	.0261562	-5.53	0.000	-.1959575	-.0933855
smsa66	.0185311	.0207149	0.89	0.371	-.0220858	.0591481
reg662	.1007678	.0361272	2.79	0.005	.0299312	.1716044
reg663	.1482588	.0352916	4.20	0.000	.0790604	.2174571
reg664	.0498971	.0419309	1.19	0.234	-.0323192	.1321134
reg665	.1462719	.0451176	3.24	0.001	.0578073	.2347365
reg666	.1629029	.0497628	3.27	0.001	.0653302	.2604757
reg667	.1345722	.0473592	2.84	0.005	.0417123	.2274321
reg668	-.083077	.0568776	-1.46	0.144	-.1946002	.0284462
reg669	.1078142	.0400844	2.69	0.007	.0292184	.1864101
v2	-.0570621	.0528071	-1.08	0.280	-.1606039	.0464798
_cons	3.666152	.8865821	4.14	0.000	1.92778	5.404524

When we add $\widehat{v_2}$ to the original equation and estimate it by OLS, the coefficient on $\widehat{v_2}$ is about –.057 with a *t* statistic of about –1.08. While the difference in the estimates of the return to education is practically large, it is not statistically significant.

Is educ endogenous? → apply the Test for endogeneity (Hausman-Wu).

(ii) We now add *nearc2* as an IV along with *nearc4*. The 2SLS estimate of β_1 is now 0.157, se($\widehat{\beta_1}$) = .052. The estimate is even larger (as the estimate obtained using one IV *nearc4*) , and statistically significant at 1%.

```
. ivregress 2sls lwage (educ=nearc2 nearc4) exper expersq black smsa south smsa66 reg662-reg669, first
```

First-stage regressions

```
                                                Number of obs =  3,010
                                                F(16, 2993)   = 170.99
                                                Prob > F      = 0.0000
                                                R-squared     = 0.4776
                                                Adj R-squared = 0.4748
                                                Root MSE      = 1.9400
```

educ	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
exper	-.4122915	.0336914	-12.24	0.000	-.4783521	-.3462309
expersq	.0008479	.00165	0.51	0.607	-.0023874	.0040832
black	-.9451729	.0939073	-10.06	0.000	-1.129302	-.7610434
smsa	.4013708	.1047858	3.83	0.000	.1959113	.6068303
south	-.0419115	.1355316	-0.31	0.757	-.3076561	.2238331
smsa66	.0000782	.1069445	0.00	0.999	-.2096139	.2097704
reg662	-.1002481	.1875618	-0.53	0.593	-.4680113	.2675151
reg663	-.0214286	.1833737	-0.12	0.907	-.3809798	.3381226
reg664	.1310678	.2173736	0.60	0.547	-.295149	.5572847
reg665	-.2683558	.2183813	-1.23	0.219	-.6965485	.1598369
reg666	-.3334436	.2377938	-1.40	0.161	-.7996995	.1328123
reg667	-.2087488	.2343833	-0.89	0.373	-.6683174	.2508198
reg668	.5507871	.2679423	2.06	0.040	.0254175	1.076157
reg669	.1687829	.2040832	0.83	0.408	-.2313747	.5689405
nearc2	.1229986	.0774256	1.59	0.112	-.0288142	.2748114
nearc4	.3205819	.0878425	3.65	0.000	.148344	.4928197
_cons	16.60428	.2415174	68.75	0.000	16.13072	17.07783

```
Instrumental variables 2SLS regression          Number of obs   =      3,010
                                                Wald chi2(15)   =     709.89
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.1702
                                                Root MSE        =      .4042
```

lwage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
educ	.1570594	.0524383	3.00	0.003	.0542822	.2598366
exper	.1188149	.0227454	5.22	0.000	.0742348	.163395
expersq	-.0023565	.0003466	-6.80	0.000	-.0030358	-.0016772
black	-.1232778	.0520112	-2.37	0.018	-.225218	-.0213376
smsa	.100753	.0314355	3.21	0.001	.0391406	.1623654
south	-.1431945	.0283691	-5.05	0.000	-.1987968	-.0875921
smsa66	.0150626	.0222765	0.68	0.499	-.0285986	.0587238
reg662	.1027473	.0391861	2.62	0.009	.0259441	.1795506
reg663	.1499316	.0382896	3.92	0.000	.0748853	.2249779
reg664	.0475676	.0454799	1.05	0.296	-.0415714	.1367066
reg665	.1544801	.0484336	3.19	0.001	.0595521	.2494082
reg666	.1729728	.0532743	3.25	0.001	.0685572	.2773884
reg667	.1420356	.0509858	2.79	0.005	.0421052	.2419659
reg668	-.0950611	.0608178	-1.56	0.118	-.2142617	.0241396
reg669	.102976	.0433068	2.38	0.017	.0180962	.1878558
_cons	3.236711	.8825567	3.67	0.000	1.506931	4.96649

```
Endogenous: educ
Exogenous:  exper expersq black smsa south smsa66 reg662 reg663 reg664
            reg665 reg666 reg667 reg668 reg669 nearc2 nearc4
```

Additional Material: do the potential instruments fulfill the relevance condition?

```
. reg educ nearc2 nearc4 exper expersq black smsa smsa66 south reg662-reg669

      Source |       SS           df       MS      Number of obs   =     3,010
-------------+----------------------------------   F(16, 2993)     =    170.99
       Model |  10297.1164        16  643.569774   Prob > F        =    0.0000
    Residual |  11264.9637     2,993  3.76377002   R-squared       =    0.4776
-------------+----------------------------------   Adj R-squared   =    0.4748
       Total |  21562.0801     3,009  7.16586243   Root MSE        =      1.94

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      nearc2 |   .1229986   .0774256     1.59   0.112    -.0288142    .2748114
      nearc4 |   .3205819   .0878425     3.65   0.000      .148344    .4928197
       exper |  -.4122915   .0336914   -12.24   0.000    -.4783521   -.3462309
     expersq |   .0008479     .00165     0.51   0.607    -.0023874    .0040832
       black |  -.9451729   .0939073   -10.06   0.000    -1.129302   -.7610434
        smsa |   .4013708   .1047858     3.83   0.000     .1959113    .6068303
      smsa66 |   .0000782   .1069445     0.00   0.999    -.2096139    .2097704
       south |  -.0419115   .1355316    -0.31   0.757    -.3076561    .2238331
      reg662 |  -.1002481   .1875618    -0.53   0.593    -.4680113    .2675151
      reg663 |  -.0214286   .1833737    -0.12   0.907    -.3809798    .3381226
      reg664 |   .1310678   .2173736     0.60   0.547     -.295149    .5572847
      reg665 |  -.2683558   .2183813    -1.23   0.219    -.6965485    .1598369
      reg666 |  -.3334436   .2377938    -1.40   0.161    -.7996995    .1328123
      reg667 |  -.2087488   .2343833    -0.89   0.373    -.6683174    .2508198
      reg668 |   .5507871   .2679423     2.06   0.040     .0254175    1.076157
      reg669 |   .1687829   .2040832     0.83   0.408    -.2313747    .5689405
       _cons |   16.60428   .2415174    68.75   0.000     16.13072    17.07783
```

```
. test nearc2 nearc4

 ( 1)  nearc2 = 0
 ( 2)  nearc4 = 0

       F(  2,  2993) =    7.89
            Prob > F =    0.0004
```

- The instruments fulfill the relevance requirement.
- The value of the F-statistics is 7.89, which is less than 10, the value used as a rule of thumb.

Utrecht University

iii) Test for overidentifying restriction

As we included two instruments, we need to check for overidentification.

After running 2SLS

ivregress 2sls lwage (educ=nearc2 nearc4) exper expersq black smsa south smsa66 reg662-reg669, first

Let $\hat{u}_i$ be the 2SLS residuals. We regress these on all exogenous variables, including *nearc2* and *nearc4*.

The *n*-*R*-squared statistic is (3,010)(.0004) ≈ 1.20.

Ho: no overidentification

H1: overidentification

We have indication of overidentification if :

n$* R^2 > \chi^2_{q,\alpha}$

1.20 < 3.84, there is no indication of overidentifcation

```
. predict uhat, res

. reg uhat nearc2 nearc4 exper expersq black smsa south smsa66 reg662-reg669
```

Source	SS	df	MS
Model	.203922835	16	.012745177
Residual	491.568721	2,993	.164239466
Total	491.772644	3,009	.163433913

Number of obs	=	3,010
F(16, 2993)	=	0.08
Prob > F	=	1.0000
R-squared	=	0.0004
Adj R-squared	=	-0.0049
Root MSE	=	.40526

uhat	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
nearc2	.0165189	.0161738	1.02	0.307	-.015194	.0482318
nearc4	-.0080835	.0183498	-0.44	0.660	-.044063	.0278961
exper	.0000312	.0070379	0.00	0.996	-.0137685	.0138309
expersq	-3.43e-06	.0003447	-0.01	0.992	-.0006793	.0006724
black	-.0008852	.0196167	-0.05	0.964	-.0393489	.0375784
smsa	.0006683	.0218892	0.03	0.976	-.0422511	.0435877
south	.0011448	.0283118	0.04	0.968	-.0543678	.0566574
smsa66	-.0005942	.0223401	-0.03	0.979	-.0443978	.0432093
reg662	-.0028725	.0391807	-0.07	0.942	-.0796962	.0739513
reg663	-.0000845	.0383058	-0.00	0.998	-.0751928	.0750238
reg664	.0011997	.0454082	0.03	0.979	-.0878347	.0902341
reg665	-.0006692	.0456187	-0.01	0.988	-.0901163	.088778
reg666	-.0064448	.0496738	-0.13	0.897	-.1038431	.0909536
reg667	-.0008388	.0489614	-0.02	0.986	-.0968402	.0951626
reg668	.0022079	.0559717	0.04	0.969	-.107539	.1119548
reg669	-.0061073	.0426319	-0.14	0.886	-.089698	.0774835
_cons	-.0003224	.0504517	-0.01	0.995	-.0992459	.098601

Alternative STATA command for Test for Overidentifaction Hansen J test (Sargan Test)

```
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  1.24815  (p = 0.2639)
  Basmann chi2(1)        =  1.24162  (p = 0.2652)
```

Ho: no overidentification

H1: overidentification

If $\chi_q^2 > \chi_{cv}^2$, reject Ho

1.25 < 3.84, fail to reject Ho

We conclude that there is no overidentification.

Utrecht
University

C2 Use MROZ.RAW for this exercise.

(i) Reestimate the labor supply function in Example 16.5, using log(*hours*) as the dependent variable. Compare the estimated elasticity (which is now constant) to the estimate obtained from equation (16.24) at the average hours worked.

(ii) In the labor supply equation from part (i), allow *educ* to be endogenous because of omitted ability. Use *motheduc* and *fatheduc* as IVs for *educ*. Remember, you now have two endogenous variables in the equation.

(iii) Test the overidentifying restrictions in the 2SLS estimation from part (ii). Do the IVs pass the test?

EXAMPLE 16.5 LABOR SUPPLY OF MARRIED, WORKING WOMEN

We use the data on working, married women in MROZ.RAW to estimate the labor supply equation (16.19) by 2SLS. The full set of instruments includes *educ*, *age*, *kidslt6*, *nwifeinc*, *exper*, and $exper^2$. The estimated labor supply curve is

$$\begin{aligned}\widehat{hours} &= \underset{(574.56)}{2{,}225.66} + \underset{(470.58)}{1{,}639.56}\,\log(wage) - \underset{(59.10)}{183.75}\,educ \\ &\quad - \underset{(9.38)}{7.81}\,age - \underset{(182.93)}{198.15}\,kidslt6 - \underset{(6.61)}{10.17}\,nwifeinc \\ n &= 428,\end{aligned} \quad [16.24]$$

which shows that the labor supply curve slopes upward. The estimated coefficient on log(*wage*) has the following interpretation: holding other factors fixed, $\Delta\widehat{hours} \approx 16.4(\%\Delta wage)$. We can calculate labor supply elasticities by multiplying both sides of this last equation by 100/*hours*:

$$100\cdot(\Delta\widehat{hours}/hours) \approx (1{,}640/hours)(\%\Delta wage)$$

or

$$\%\Delta\widehat{hours} \approx (1{,}640/hours)(\%\Delta wage),$$

which implies that the labor supply elasticity (with respect to wage) is simply 1,640/*hours*. [The elasticity is not constant in this model because *hours*, not log(*hours*), is the dependent variable in (16.24).] At the average hours worked, 1,303, the estimated elasticity is 1,640/1,303 ≈ 1.26, which implies a greater than 1% increase in hours worked given a 1% increase in wage. This is a large estimated elasticity. At higher hours, the elasticity will be smaller; at lower hours, such as *hours* = 800, the elasticity is over two.

For comparison, when (16.19) is estimated by OLS, the coefficient on log(*wage*) is −2.05 (se = 54.88), which implies no wage effect on hours worked. To confirm that log(*wage*) is in fact endogenous in (16.19), we can carry out the test from Section 15.5. When we add the reduced form residuals $\hat{v}_2$ to the equation and estimate by OLS, the *t* statistic on $\hat{v}_2$ is −6.61, which is very significant, and so log(*wage*) appears to be endogenous.

$$\begin{aligned}hours &= \alpha_1\log(wage) + \beta_{10} + \beta_{11}educ + \beta_{12}age + \beta_{13}kidslt6 \\ &\quad + \beta_{14}nwifeinc + u_1\end{aligned} \quad [16.19]$$

$$\begin{aligned}\log(wage) &= \alpha_2 hours + \beta_{20} + \beta_{21}educ + \beta_{22}exper \\ &\quad + \beta_{23}exper^2 + u_2.\end{aligned} \quad [16.20]$$

The wage offer equation (16.20) can also be estimated by 2SLS. The result is

$$\begin{aligned}\widehat{\log(wage)} &= -\underset{(.338)}{.656} + \underset{(.00025)}{.00013}\,hours + \underset{(.016)}{.110}\,educ \\ &\quad + \underset{(.019)}{.035}\,exper - \underset{(.00045)}{.00071}\,exper^2 \\ n &= 428.\end{aligned} \quad [16.25]$$

This differs from previous wage equations in that *hours* is included as an explanatory variable and 2SLS is used to account for endogeneity of *hours* (and we assume that *educ* and *exper* are exogenous). The coefficient on *hours* is statistically insignificant, which means that there is no evidence that the wage offer increases with hours worked. The other coefficients are similar to what we get by dropping *hours* and estimating the equation by OLS.

Utrecht University

(i). Generate first log(hours), run the 2sls regression and compare to 16.24.

```
. gen lhours= ln(hours)
(325 missing values generated)

. ivregress 2sls lhours (lwage=exper expersq) educ age kidslt6 nwifeinc, first

First-stage regressions

                                          Number of obs =      428
                                          F(6, 421)     =    13.69
                                          Prob > F      =   0.0000
                                          R-squared     =   0.1633
                                          Adj R-squared =   0.1514
                                          Root MSE      =   0.6662
```

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.1011113	.0149618	6.76	0.000	.0717023	.1305204
age	-.0025561	.005192	-0.49	0.623	-.0127615	.0076492
kidslt6	-.0532185	.0884411	-0.60	0.548	-.2270596	.1206225
nwifeinc	.00556	.0033104	1.68	0.094	-.0009469	.0120669
exper	.0418643	.0132377	3.16	0.002	.015844	.0678846
expersq	-.0007625	.0004008	-1.90	0.058	-.0015503	.0000253
_cons	-.4471607	.2852028	-1.57	0.118	-1.00776	.1134381

```
Instrumental variables 2SLS regression            Number of obs   =        428
                                                  Wald chi2(5)    =      24.39
                                                  Prob > chi2     =     0.0002
                                                  R-squared       =          .
                                                  Root MSE        =     1.6125
```

lhours	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
lwage	1.994349	.5603396	3.56	0.000	.8961034	3.092594
educ	-.2354609	.0703733	-3.35	0.001	-.37339	-.0975318
age	-.0135248	.0111669	-1.21	0.226	-.0354115	.008362
kidslt6	-.4654385	.2178235	-2.14	0.033	-.8923648	-.0385123
nwifeinc	-.0139044	.0078765	-1.77	0.078	-.0293421	.0015333
_cons	8.370233	.6841642	12.23	0.000	7.029296	9.71117

```
Endogenous:  lwage
Exogenous:   educ age kidslt6 nwifeinc exper expersq
```

$$
\begin{aligned}
\widehat{hours} &= \underset{(574.56)}{2{,}225.66} + \underset{(470.58)}{1{,}639.56} \log(wage) - \underset{(59.10)}{183.75}\, educ \\
&\quad - \underset{(9.38)}{7.81}\, age - \underset{(182.93)}{198.15}\, kidslt6 - \underset{(6.61)}{10.17}\, nwifeinc \\
n &= 428,
\end{aligned}
\qquad [16.24]
$$

We estimate a constant elasticity version of the labor supply equation (naturally, only for *hours* > 0), again by 2SLS. We get

$$\log(hours) = 8.37 + 1.99 \log(wage) - .235\, educ - .014\, age - .465\, kidslt6 - .014\, nwifeinc$$

which implies a labor supply elasticity of 1.99. This is much higher than the elasticity 1.26 (E_L =1,640/1,330 = 1.26) we obtained from equation (16.24) at the mean value of hours (1303). The higher elasticity suggests that workers are more likely to increase their working hours when wages arise.

Additional Material: Test the Relevance of the excluded instruments and the overidentification restriction.

Estimate the first stage regression.

```
. reg lwage exper expersq educ age kidslt6 nwifeinc

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(6, 421)       =     13.69
       Model |  36.4697152         6  6.07828587   Prob > F        =    0.0000
    Residual |  186.857726       421  .443842579   R-squared       =    0.1633
-------------+----------------------------------   Adj R-squared   =    0.1514
       Total |  223.327441       427  .523015084   Root MSE        =    .66622

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       exper |   .0418643   .0132377     3.16   0.002      .015844    .0678846
     expersq |  -.0007625   .0004008    -1.90   0.058    -.0015503    .0000253
        educ |   .1011113   .0149618     6.76   0.000     .0717023    .1305204
         age |  -.0025561    .005192    -0.49   0.623    -.0127615    .0076492
     kidslt6 |  -.0532185   .0884411    -0.60   0.548    -.2270596    .1206225
    nwifeinc |     .00556   .0033104     1.68   0.094    -.0009469    .0120669
       _cons |  -.4471607   .2852028    -1.57   0.118     -1.00776    .1134381
------------------------------------------------------------------------------
```

Do the exclusion test:

```
. test exper expersq

 ( 1)  exper = 0
 ( 2)  expersq = 0

       F(  2,   421) =    9.33
            Prob > F =    0.0001
```

They are statistically significant, and the relevance condition is fulfilled.
The Fstatistics is greater than the critical value (3.00), we reject Ho.
The F-test is less than 10.

```
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .068537  (p = 0.7935)
  Basmann chi2(1)        =  .067426  (p = 0.7951)
```

Ho: no overidentification
H1: overidentification
If $\chi_q^2 > \chi_{cv}^2$, reject Ho
0.068 < 3.84, there is no indication of overidentifcation

The overidentifying assumptions cannot be rejected at any level of significance.

(ii) Now we estimate the equation by 2SLS but allow log(*wage*) and *educ* to both be endogenous. The full list of instrumental variables is *age*, *kidslt6*, *nwifeinc*, *exper*, *exper2*, *motheduc*, and *fatheduc*.

```
. . ivregress 2sls lhours (lwage educ=exper expersq motheduc fatheduc) age kidslt6 nwifeinc, first

First-stage regressions

                                                Number of obs =     428
                                                F(7, 420)     =    5.00
                                                Prob > F      = 0.0000
                                                R-squared     = 0.0770
                                                Adj R-squared = 0.0616
                                                Root MSE      = 0.7006
```

lwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	-.0042493	.0055751	-0.76	0.446	-.0152078	.0067093
kidslt6	.0201473	.0923331	0.22	0.827	-.1613451	.2016398
nwifeinc	.0116132	.0033573	3.46	0.001	.005014	.0182124
exper	.0485847	.0138808	3.50	0.001	.0213002	.0758691
expersq	-.0008773	.0004214	-2.08	0.038	-.0017056	-.000049
motheduc	.0009898	.0125966	0.08	0.937	-.0237704	.02575
fatheduc	.0134516	.0116927	1.15	0.251	-.0095319	.0364351
_cons	.5879528	.2761049	2.13	0.034	.0452333	1.130672

```
Number of obs =     428
F(7, 420)     =   23.38
Prob > F      = 0.0000
R-squared     = 0.2804
Adj R-squared = 0.2684
Root MSE      = 1.9548
```

educ	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	.0001603	.0155558	0.01	0.992	-.0304167	.0307373
kidslt6	.7347522	.2576312	2.85	0.005	.2283449	1.241159
nwifeinc	.0528863	.0093676	5.65	0.000	.0344731	.0712995
exper	.0603897	.0387307	1.56	0.120	-.0157404	.1365198
expersq	-.0009699	.0011758	-0.82	0.410	-.0032811	.0013413
motheduc	.1552934	.0351475	4.42	0.000	.0862065	.2243802
fatheduc	.166468	.0326255	5.10	0.000	.1023385	.2305976
_cons	8.013848	.7703985	10.40	0.000	6.499531	9.528165

```
Instrumental variables 2SLS regression          Number of obs   =        428
                                                Wald chi2(5)    =      26.28
                                                Prob > chi2     =     0.0001
                                                R-squared       =          .
                                                Root MSE        =     1.5247
```

lhours	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
lwage	1.810915	.4942677	3.66	0.000	.8421684	2.779662
educ	-.1286057	.0868182	-1.48	0.139	-.2987662	.0415548
age	-.0116012	.0105112	-1.10	0.270	-.0322027	.0090003
kidslt6	-.5431861	.2098478	-2.59	0.010	-.9544802	-.1318919
nwifeinc	-.0189058	.0087217	-2.17	0.030	-.036	-.0018116
_cons	7.260764	1.012221	7.17	0.000	5.276847	9.244681

```
Endogenous:  lwage educ
Exogenous:   age kidslt6 nwifeinc exper expersq motheduc fatheduc
```

The biggest effect is to reduce the size of the coefficient on *educ* as well as its statistical significance. The labor supply elasticity is only moderately smaller.

Additional Material: Check the relevance of the instruments.

```
. reg lwage exper expersq fatheduc motheduc age kidslt6 nwifeinc

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(7, 420)       =      5.00
       Model |  17.1897236         7  2.4556748   Prob > F        =    0.0000
    Residual |  206.137717       420  .490804089   R-squared       =    0.0770
-------------+----------------------------------   Adj R-squared   =    0.0616
       Total |  223.327441       427  .523015084   Root MSE        =    .70057

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       exper |   .0485847   .0138808     3.50   0.001     .0213002    .0758691
     expersq |  -.0008773   .0004214    -2.08   0.038    -.0017056   -.000049
    fatheduc |   .0134516   .0116927     1.15   0.251    -.0095319    .0364351
    motheduc |   .0009898   .0125966     0.08   0.937    -.0237704      .02575
         age |  -.0042493   .0055751    -0.76   0.446    -.0152078    .0067093
     kidslt6 |   .0201473   .0923331     0.22   0.827    -.1613451    .2016398
    nwifeinc |   .0116132   .0033573     3.46   0.001      .005014    .0182124
       _cons |   .5879528   .2761049     2.13   0.034     .0452333    1.130672
------------------------------------------------------------------------------
```

```
. test exper expersq fatheduc motheduc

 ( 1)  exper = 0
 ( 2)  expersq = 0
 ( 3)  fatheduc = 0
 ( 4)  motheduc = 0

       F(  4,   420) =    6.29
            Prob > F =    0.0001
```

iii) Test for the overidentification assumptions:

```
estat overid

Tests of overidentifying restrictions:

Sargan (score) chi2(2) =  .446375  (p = 0.8000)
Basmann chi2(2)        =  .438489  (p = 0.8031)
```

There are 4 excluded instruments and 2 endogenous variables.
Ho: no overidentification
H1: overidentification
If $\chi_2^2 > \chi_{cv}^2$, reject Ho.
$0.45 < 5.99$, fail to reject Ho.
There is no overidentification.

We can also do it like this:

```
. predict uhat, res
(325 missing values generated)

.  reg uhat exper expersq age kidslt6 nwifeinc motheduc fatheduc
```

Source	SS	df	MS
Model	1.03769873	7	.148242676
Residual	993.943409	420	2.36653193
Total	994.981108	427	2.33016653

Number of obs	=	428
F(7, 420)	=	0.06
Prob > F	=	0.9996
R-squared	=	0.0010
Adj R-squared	=	-0.0156
Root MSE	=	1.5384

uhat	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
exper	-.0043184	.0304801	-0.14	0.887	-.0642309	.0555941
expersq	.0001953	.0009253	0.21	0.833	-.0016235	.0020142
age	-.0005731	.0122421	-0.05	0.963	-.0246365	.0234902
kidslt6	.0031535	.2027493	0.02	0.988	-.3953762	.4016832
nwifeinc	.0005676	.0073721	0.08	0.939	-.0139232	.0150583
motheduc	.0144205	.0276602	0.52	0.602	-.0399491	.0687901
fatheduc	-.0131371	.0256754	-0.51	0.609	-.0636055	.0373312
_cons	.0041712	.6062841	0.01	0.995	-1.187558	1.1959

After obtaining the 2SLS residuals, $\hat{u}_i$ from the estimation in part (ii), we regress these on *age*, *kidslt6*, *nwifeinc*, *exper*, exper2, *motheduc*, and *fatheduc*.
The *n*-*R*-squared statistic is 428(.0010) = .428.
Ho: no overidentification
H1: overidentification
0.428<5.99, we fail to reject Ho.
There is no overidentification.