Problem Set 2 Solutions
EE426 Semester 2/2014

(1.1) Substitute the reduced form into the structural equation to get

$y_1 = \beta_0 + \beta_1(\pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2) + \beta_2 z_1 + u_1$

$y_1 = \beta_0 + \beta_1\pi_0 + \beta_1\pi_1 z_1 + \beta_1\pi_2 z_2 + \beta_1 v_2 + \beta_2 z_1 + u_1$

$y_1 = (\beta_0 + \beta_1\pi_0) + (\beta_1\pi_1 + \beta_2)z_1 + \beta_1\pi_2 z_2 + (\beta_1 v_2 + u_1)$

So, we have $\alpha_0 = \beta_0 + \beta_1\pi_0, \alpha_1 = \beta_1\pi_1 + \beta_2$, and $\alpha_2 = \beta_1\pi_2$.

(1.2) $v_1 = \beta_1 v_2 + u_1$

(1.3) To consistently estimate the α_j, we need $E(v_1) = 0$ and no correlations between v_1 and z_1, v_1 and z_2. That's is cov(v_1, z_1) = 0 and cov(v_1, z_2) = 0

(2.1) The results are reported below.

```
-----------------------------
                      reg21
                       b/se
-----------------------------
educ                -0.091***
                   (0.0059)
age                  0.332***
                   (0.0165)
agesq               -0.003***
                   (0.0003)
_cons               -4.138***
                   (0.2406)
-----------------------------
r2                    0.569
N                      4361
-----------------------------
*** P<0.01, ** P<0.05, and * P<0.10.
```

The coefficient on year of education is -0.091. Holding age fixed, another year of education will lower fertility for 0.091 children. In other words, for a group of 100 women, another year of education in this group will predict to have nine fewer children on average.

(2.2) The reduced form (first stage) for educ is

```
-----------------------------
                      reg22
                       b/se
-----------------------------
frsthalf            -0.852***
                   (0.1128)
age                 -0.108**
                   (0.0420)
```

```
agesq              -0.001
                  (0.0007)
_cons               9.693***
                  (0.5981)
----------------------------
r2                  0.108
N                   4361
----------------------------
*** P<0.01, ** P<0.05, and * P<0.10.
```

The coefficient on frsthalf is -0.852 and statistically significantly different from zero at 0.01 (s.e. = 0.1128, t-statistic is -7.55). Women born in the first half of the year are predicted to have almost one year less education. So, we have frsthalf strongly predict educ, implying that we can use frsthalf as our IV candidate for educ.

(2.3) The estimates are reported below, comparing OLS and IV (using frsthalf as an IV for educ).

```
--------------------------------------------
                     OLS              IV
                     b/se            b/se
--------------------------------------------
educ                -0.091***       -0.171***
                   (0.0059)        (0.0532)
age                  0.332***        0.324***
                   (0.0165)        (0.0179)
agesq               -0.003***       -0.003***
                   (0.0003)        (0.0003)
_cons               -4.138***       -3.388***
                   (0.2406)        (0.5479)
--------------------------------------------
r2                   0.569           0.550
N                    4361            4361
--------------------------------------------
*** P<0.01, ** P<0.05, and * P<0.10.
```

The estimated effect of education on fertility is much bigger. For a group of 100 women, another year of education will predict to have seventeen fewer children on average. We can see that the estimated standard error is also bigger (0.0532 for IV compared to 0.0059 for OLS). This produces a fairly wide 95% confidence interval for the coefficient on education.

(2.4) Which model should we rely on?
>> Test for endogeneity

Method #1 predict vhat (residuals) from first stage regression, then put it into the structural equation. Null hypothesis is that the coefficient on vhat is 0, meaning that our education variable is exogenous. If we reject H_0, our education variable is endogenous, and hence using IV is more suitable.

```
reg children educ age agesq vhat
```

```
      Source |       SS       df       MS              Number of obs =    4361
-------------+------------------------------           F(  4,  4356) = 1437.49
       Model |  12248.2426     4  3062.06066           Prob > F      =  0.0000
    Residual |   9278.9337  4356  2.13015007           R-squared     =  0.5690
-------------+------------------------------           Adj R-squared =  0.5686
       Total |  21527.1763  4360  4.93742577           Root MSE      =  1.4595

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.1714989   .0520663    -3.29   0.001    -.2735754   -.0694224
         age |   .3236052   .0174857    18.51   0.000     .2893243    .3578861
       agesq |  -.0026723   .0002738    -9.76   0.000    -.0032091   -.0021354
        vhat |   .0819832   .0524061     1.56   0.118    -.0207595    .1847259
       _cons |  -3.387805   .5366747    -6.31   0.000    -4.439961    -2.33565
------------------------------------------------------------------------------
```

Here, we fail to reject H_0 at 10% level of significance, hence OLS estimator is more preferable.

Method #2: Hausman test IV vs. OLS
Note: reg23 = IV, reg21 = OLS

```
hausman reg23 reg21, constant sigmamore

                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     reg23        reg21        Difference          S.E.
-------------+------------------------------------------------------------
        educ |   -.1714989    -.0905755       -.0809235        .0517373
         age |    .3236052     .3324486       -.0088434        .0056539
       agesq |   -.0026723    -.0026308       -.0000415        .0000265
       _cons |   -3.387805    -4.138307        .7505013        .4798227
------------------------------------------------------------------------------
                     b = consistent under Ho and Ha; obtained from ivregress
          B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        2.45
                Prob>chi2 =      0.1178
                (V_b-V_B is not positive definite)
```

Here, we cannot reject H_0 at 10% as well. The difference in coefficients across the two models is not systematic and not significant, implying that OLS is both consistent and efficient estimator, or OLS estimator is more preferable.

(3.1) First stage regression:
$educ = 16.6383 + 0.3199 nearc4 - 0.4125 exper + 0.0009 expersq - 0.9355 black + 0.4022 smsa - 0.0516 south + 0.0255 smsa66 + D_{reg}*\beta$

The t-statistic for nearc4 is 0.3199/0.0879 = 3.64, which confirms a statistical significance of using *nearc4* in predicting *educ*. Alternatively, if we do F-test for the coefficient on nearc4, we have F = 13.26 with p-value = 0.0003. We can use nearc4 as an IV for education.

(3.2) The OLS and IV estimations are reported below.

	ols	iv
	b/se	b/se
educ	0.0747***	0.1315**
	(0.0035)	(0.0548)
exper	0.0848***	0.1083***
	(0.0066)	(0.0236)
expersq	−0.0023***	−0.0023***
	(0.0003)	(0.0003)
black	−0.1990***	−0.1468***
	(0.0182)	(0.0538)
smsa	0.1364***	0.1118***
	(0.0201)	(0.0316)
south	−0.1480***	−0.1447***
	(0.0260)	(0.0272)
smsa66	0.0262	0.0185
	(0.0194)	(0.0216)
reg662	0.0964***	0.1008***
	(0.0359)	(0.0376)
reg663	0.1445***	0.1483***
	(0.0351)	(0.0367)
reg664	0.0551	0.0499
	(0.0417)	(0.0436)
reg665	0.1280***	0.1463***
	(0.0418)	(0.0469)
reg666	0.1405***	0.1629***
	(0.0452)	(0.0518)
reg667	0.1180***	0.1346***
	(0.0448)	(0.0493)
reg668	−0.0564	−0.0831
	(0.0513)	(0.0592)
reg669	0.1186***	0.1078***
	(0.0388)	(0.0417)
_cons	4.6208***	3.6662***
	(0.0742)	(0.9224)
r2	0.300	0.238
N	3010	3010

*** P<0.01, ** P<0.05, and * P<0.10.

The estimated return to education for OLS regression is about 7.47% while the estimated return for IV regression is 13.15%, almost two times larger. All these coefficients on education are statistically significant at 1% for OLS and 5% for IV. The standard error of IV estimate is over 18 times larger than the OLS standard error. This results in a wider range of 95% confidence interval for the IV estimates, a cost to get a consistent estimator when we believe that *educ* is endogenous.

(3.3) After estimating the reduced form in (3.1), we create residuals, $\widehat{v_2}$
Run the structural/original equation by also adding $\widehat{v_2}$ as one of the explanatory variables. (See STATA code). The coefficient on $\widehat{v_2}$ is -0.057 with t-statistic = -1.08 and p-value = 0.280. Hence, although we see a large difference in the estimates of return to

education, it is not statistically significant. In other words, there is not enough evidence for *educ* being endogenous in the structural equation.

(3.4) IV1 reports the results when only nearc4 is an IV, while IV2 reports the results when both nearc4 and nearc2 are used as IVs.

	iv1	iv2
	b/se	b/se
educ	0.1315**	0.1571***
	(0.0548)	(0.0524)
exper	0.1083***	0.1188***
	(0.0236)	(0.0227)
expersq	−0.0023***	−0.0024***
	(0.0003)	(0.0003)
black	−0.1468***	−0.1233**
	(0.0538)	(0.0520)
smsa	0.1118***	0.1008***
	(0.0316)	(0.0314)
south	−0.1447***	−0.1432***
	(0.0272)	(0.0284)
smsa66	0.0185	0.0151
	(0.0216)	(0.0223)
reg662	0.1008***	0.1027***
	(0.0376)	(0.0392)
…		
reg669	0.1078***	0.1030**
	(0.0417)	(0.0433)
_cons	3.6662***	3.2367***
	(0.9224)	(0.8826)
r2	0.238	0.170
N	3010	3010

*** P<0.01, ** P<0.05, and * P<0.10.

The estimated return to education is now 15.71%, which is larger than the model using only *nearc4* and is now statistically significant at 0.01 level. However, the standard error for educ coefficient estimate in IV2 is somewhat smaller. Note that in the reduced form, nearc2 is not statistically significantly determining educ (p-value = 0.112).

(3.5) Testing for overidentifying restriction
From (3.4 – IV2), we create $\hat{u}$ from 2SLS. Then regress $\hat{u}$ on all exogenous variables, including *nearc4* and *nearc2*. Then, calculate the $n*R^2 = (3,010)*(0.0004) = 1.204$. Degree of freedom (q) is # of IVs from outside the model - #endogenous variables = 2-1 = 1. Our null hypothesis is that all IVs are uncorrelated with *u*. Here, we have $n*R^2 < \chi_1^2$ (either at 0.10 or 0.05 significance level). Therefore, we cannot reject our null hypothesis, meaning that the overidentifying restriction is not rejected (all instruments are exogenous).

```
cd "/Users/namtipka/Documents/BE courses/EE426/Problem Set/"
```

STATA commands

```
use "FERTIL2.dta"

*2.1
reg children educ age agesq
est sto reg21

estout reg21, cells(b(star fmt(4)) se(par fmt(4))) stats(r2 N, fmt(3 0))
starlevels(* 0.10 ** 0.05 *** 0.01)


*2.2
reg educ frsthalf age agesq
est sto reg22

predict vhat, resid

test frsthalf
/*  t-stat: -7.55   0.000
 ( 1)  frsthalf = 0

       F(  1,  4357) =   57.06
            Prob > F =    0.0000 */

estout reg22, cells(b(star fmt(4)) se(par fmt(4))) stats(r2 N, fmt(3 0))
starlevels(* 0.10 ** 0.05 *** 0.01)


*2.3
ivregress 2sls children (educ = frsthalf) age agesq
est sto reg23


estout reg21 reg23, cells(b(star fmt(4)) se(par fmt(4))) stats(r2 N, fmt(3
0)) starlevels(* 0.10 ** 0.05 *** 0.01)


*2.4
reg children educ age agesq vhat

*or
hausman reg23 reg21, constant sigmamore




use "CARD.dta"
*3.1
reg educ nearc4 exper expersq black smsa south smsa66 reg662- reg669

test nearc4
/*  t-stat: 3.64   0.000
( 1)  nearc4 = 0

       F(  1,  2994) =   13.26
            Prob > F =    0.0003  */

predict v2_hat, resid
```

```
*3.2
reg lwage educ exper expersq black smsa south smsa66 reg662- reg669
est sto ols

ivregress 2sls lwage (educ = nearc4) exper expersq black smsa south smsa66
reg662- reg669
est sto iv

estout ols iv, cells(b(star fmt(4)) se(par fmt(4))) stats(r2 N, fmt(3 0))
starlevels(* 0.10 ** 0.05 *** 0.01)


*3.3 do endogeneity testing
reg lwage educ exper expersq black smsa south smsa66 reg662- reg669 v2_hat

*     v2_hat |  -.0570621   .0528071   t = -1.08   0.280

test v2_hat
/*       F(  1,  2993) =    1.17
             Prob > F =    0.2800 */

*the coefficient on v^2 is about −.057 with a t statistic of about −1.08.
*Therefore, while the difference in the estimates of the return to
education is practically large, it is not statistically significant.


*3.4

*reduced form
reg educ nearc4 nearc2 exper expersq black smsa south smsa66 reg662-
reg669
*nearc2 is not significant (p-value = 0.112)
est sto red1

ivregress 2sls lwage (educ = nearc4 nearc2) exper expersq black smsa south
smsa66 reg662- reg669
est sto iv2

predict u_hat, resid

estout iv iv2, cells(b(star fmt(4)) se(par fmt(4))) stats(r2 N, fmt(3 0))
starlevels(* 0.10 ** 0.05 *** 0.01)

*3.5 overidentification test
reg u_hat nearc4 nearc2 exper expersq black smsa south smsa66 reg662-
reg669

*get R-sq
*Then, n*Rsq = (3,010)(.0004) ≈ 1.20
*p-value = P( ?2 >1.20) ≈ .55 >> so the overidentifying restriction is not
rejected
```