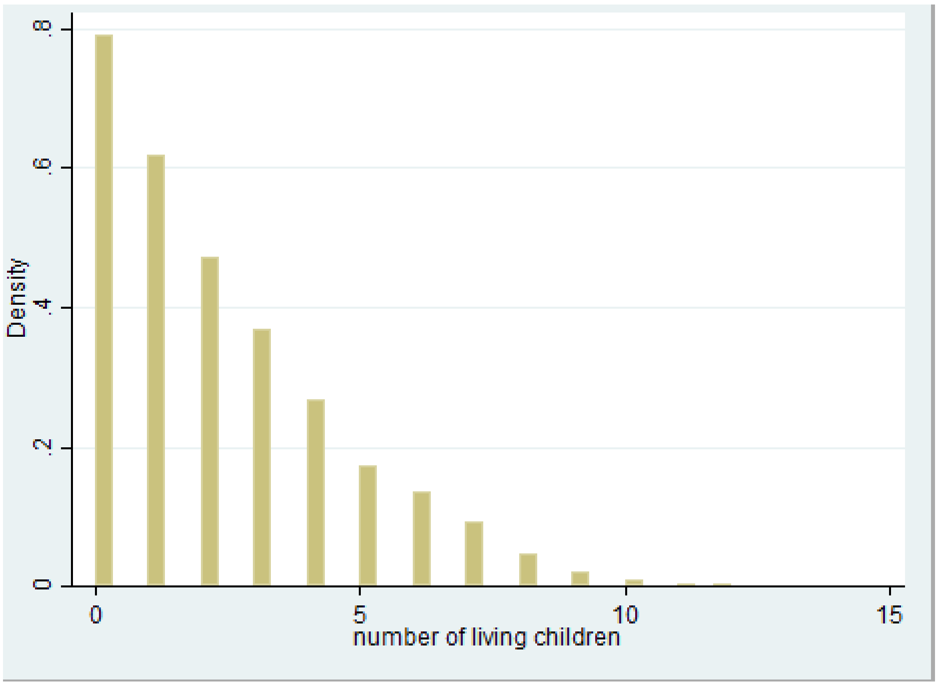
Assignment 2

SID:470518795

Part A: Descriptive Statistics for the Sample

(1)

Summary	AVG	STD	MIN	MAX	MEDIAN	10TH	25TH	75TH	90TH
Children	2.26	2.22	0	12	2	0	0	4	6



In my view, the sample average number of children is not a good measure of typical number of children. By the natural property of children, the number of children should be a positive integer or zero. (At least now)

$$P(number\ of\ children = mean(2.26)) = 0$$

Therefore, the sample average of children cannot represent number of children. Mode would be a good measure, it is most likely that the number of children is equal to mode.

It seems the sample distribution is approximately normal. By definition of normal distribution, mean=median=mode (not in our case). The normal distribution should not be skewed. Our histogram indicates a right-skewed sample distribution and not symmetric.

(2)

Summary	AVG	STD	MIN	MAX	MEDIAN
educ	5.87	3.91	0	20	7

From the summary statistics, we can see that mean is less than median. The sample distribution of educ would be left-skewed. The difference indicates that there are quite few people getting education below than average.

$Fraction\ of\ noeduc = \frac{number\ of\ non\ education\ in\ sample}{number\ of\ observations}$ =0.20375

(3)

```
. summarize age agefbrth urban tv electric frsthalf

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      4,000    27.39075    8.712321         15         49
    agefbrth |      2,989     19.0087    3.110406         10         38
       urban |      4,000      .51675    .4997818          0          1
          tv |      4,000      .09275     .290118          0          1
    electric |      4,000      .13975    .3467708          0          1
-------------+---------------------------------------------------------
    frsthalf |      4,000      .53975    .4984797          0          1

.
```

The number of observation of agefbrth is quite unusual. It probably is missing data issue.

Part B: Multiple Regression Model – Estimation and Testing

(4)

On average, ceteris paribus, an increase in one year of education is associated with a $a\ increase\ in\ \beta 1$ number of children.

(5)

$\widehat{children} = -4.131478 - .0893658 * educ + 0.3321312age - .0026402age^2$ EQ1

(.2500609) (.00623) (.3321312) (-.0026402)

$n = 4000 \quad R^2 = 0.5679$

(6)

$$H_0: B_{EDUC} = 0 \ \ VS \ \ H_a: B_{EDUC} \neq 0$$

$$Calculating\ t\ statistics\text{:}\ \ t = \frac{-.0893658 - 0}{0.00623} = -14.34$$

Rejection Rule: reject H_0 in favor of H_a when the absolute value of t is greater than c (critical value) and c the critical value for the t distribution with 3996=4000-4 df and a 1% significance level. After consulting a t table, we find c=2.576

Decision: since absolute value of t is more than critical value, we have enough evidence to reject the null hypothesis in favour of the alternative. We have sufficient evidence that the year of education will affect the number of children at 1% significance level.

(7)

The multiple regression does not provides a causal relationship. As listed in dataset, there are so many other factors which have not yet accounted for which affect and help explain variations in children. Not including them in the specification of our model, implies that they are in the error term. Hence, there is potential endogeneity concerns. It is a violation of zero conditional mean independence. Therefore, we cannot say there is a causal relationship. Our estimate is biased.

(8)

It is expected that the correlation between family income and year of educations is positive. The reason behind this is that people getting more education would have more skills and training, they would be more competitive in job market. Therefore, they are more likely to get higher wage.

$$True\ model\text{:}\ children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + \beta_4 income + u$$

$$We\ estimate\text{:}\ children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + v$$

$$v = \beta_4 income + u$$

There is a positive correlation between income and education. $Cor(income, educ) > 0$

$$income = \gamma_0 + \gamma_1 educ + v \quad where\ \ \gamma_1 > 0$$

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + \beta_4(\gamma_0 + \gamma_1 educ + v) + u$$

By rearrange the equation,

$$children = (\beta_0 + \beta_4\gamma_0) + (\beta_1 + \beta_4\gamma_1) educ + \beta_2 age + \beta_3 age^2 + (\beta_4 v + u)$$

Conclusion: All estimated coefficient are biased, intercept and slope.

$\gamma_1 > 0$ and $\beta_4 < 0, the\ product\ of\ them\ would\ be\ negative. \beta_1\ would\ be\ under-restimated\ (downward\ bias)$

It will look as if people with many years of education would have less children, but this is partly due to the effect of income - the fact that people with more income are also more educated.

Omitted variable would violate Assumption 4, our estimates are biased and inconsistent. Our linear regression will no longer be BLUE. Our inference I.e. t-statistics and CI would be invalid.

(9)

$$\widehat{children} = -4.217031 - 0.071652 * educ + 0.339836age - .002729age^2 \; - 0.203995urban$$

(0.25) (0.07) (0.34) (-0.0027) (-0.204)

$$-0.25055electric - 0.171123tv$$

(-0.251) (-0.171)

$n = 4000, R^2$ =0.5735 EQ2

(10)

The estimated coefficient in EQ2 is larger than it in EQ1. This indicates that smaller effect. This conform the Question 8. Since there is a negative bias, by using proxy variable, we diminish the bias. When the negative bias vanishes, there would be smaller effect. It is consistent with our expectation.

(11)

$$H_0: B_{urban} = B_{electric} = B_{tv} = 0 \quad VS \;\; H_a: H_0 \; is \; false.$$

Unrestricted model: EQ2

Restricted model: EQ1

SSR-UR= 8394.79318

SSR-R=8506.07115

$$F = \frac{\frac{SSR_r - SSR_{UR}}{q}}{\frac{SSR_{UR}}{n-k-1}} = 17.64$$

Rejection Rule: reject H_0 in favor of H_a when the F is greater than c (critical value) and c the critical value for the F distribution with 3993=4000-7 df and q=3. After consulting a F table, we find c=2.80

Decision: since F is more than critical value, we have enough evidence to reject the null hypothesis in favour of the alternative. We have sufficient evidence that these three variables are jointly significant at 1% significance level.

Part C: Check for heteroscedasticity

(12)

Give the regression,

$$\hat{u}^2 = \delta_0 + \delta_1 educ + \delta_2 age + \delta_3 age^2 + \delta_4 urban + \delta_5 electric + \delta_6 tv$$

Null hypothesis: $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$ i. e. homoscedasticity

Alternative hypothesis: $null\ hypothesis\ is\ false\ \ i.e. heteroskedasticity$

```
reg uhat2 educ age age2 urban electric tv
```

Source	SS	df	MS
Model	17124.9546	6	2854.15911
Residual	51554.6904	3,993	12.9112673
Total	68679.6451	3,999	17.1742048

Number of obs	=	4,000
F(6, 3993)	=	221.06
Prob > F	=	0.0000
R-squared	=	0.2493
Adj R-squared	=	0.2482
Root MSE	=	3.5932

uhat2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.1493257	.0166395	-8.97	0.000	-.1819484	-.116703
age	-.0751198	.0424439	-1.77	0.077	-.1583335	.0080938
age2	.004623	.0006986	6.62	0.000	.0032533	.0059926
urban	-.0756677	.1203375	-0.63	0.530	-.3115964	.1602611
electric	.1862056	.1993624	0.93	0.350	-.204656	.5770672
tv	-.2996352	.2364717	-1.27	0.205	-.7632517	.1639813
_cons	1.255015	.6183111	2.03	0.042	.0427799	2.46725

By looking at the table,

$$F(6{,}3993) = 221.06$$

Rejection Rule: reject H_0 in favor of H_a when the F is greater than c (critical value) and c the critical value for the F distribution with 3993=4000-7 df and q=6. After consulting a F table, we find c=2.10

Decision: since F is more than critical value, we have enough evidence to reject the null hypothesis in favour of the alternative. We have sufficient evidence that there is a heteroscedasticity issue in our original EQ2 at a 5% significant level.

(13)

$$\widehat{children} = -4.217031 - 0.071652 * educ + 0.339836age - .002729age^2 - 0.203995urban$$

(0.253) (.0067) (.0199962) (.0003658) (.0475067)

$$-0.25055electric - 0.171123tv$$

(.0790036) (.0878761)

$n = 4000, R^2 = 0.5735$

Part D: Instrumental Variables

(14)

Although there are three proxies, it is sure that there exists some income variable cannot be explained by these three proxy variables. These unexplained data would still turn to error term. The income is correlated with educ. Therefore, $cor(educ, u) \neq 0$. This is the week condition compared to independence. So, the correlation is not equal to zero indicating that $E(u|educ) \neq 0$.

Since we know the correlation between educ and income are positive, therefore, the correlation between educ and error term is positive, as we show, there will be a downward bias on our estimated coefficient if we use OLS.

(15)

Firstly, we need instrument relevance, i.e. our instrument variable should be highly correlated to endogenous variable. $Cor(educ, frsthalf) \neq 0$

Secondly, we need instrument validity, i.e. our instrument variable should be uncorrelated with error term. $Cor(frsthalf, u) = 0$

In that case, our IV estimator will be consistent. We cannot test for condition (2) (Instrument Validity), we only can use economic theory and intuition to decide if assuming that is reasonable.

However, we can test for first condition.

If $Cor(educ, frsthalf) \neq 0$ holds,

Run a regression, $educ = \gamma_0 + \gamma_1 frsthalf + \gamma_2 age + \gamma_3 age^2 + \gamma_4 urban + \gamma_5 electric + \gamma_6 tv + v$

$$H0{:}\, \gamma_1 = 0 \;vs\; H1{:}\, \gamma_1 \neq 0$$

For instrument relevance, we want γ_1 to be large and highly significant.

(16)

On average, if women born in the first six months of the year, the estimated educ=5.45. If not, educ=6.37. The difference is probably because of compulsory school attendance laws.

```
sum educ if frsthalf==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	2,159	5.450672	3.903878	0	18

```
nd of do-file

do "C:\Users\yjin5959\AppData\Local\Temp\39\STD00000000.tmp"

sum educ if frsthalf==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	1,841	6.369364	3.855602	0	20

```
nd of do-file
```

(17)

$\widehat{educ} = 9.101798 - .6609955 frsthalf - .1322433 age - .0002166\, age^2 + .8143062 urban + 1.911368 electric + 2.606173 tv$

n= 4000 R^2=0.2436

$$H_0: \gamma_{frsthalf} = 0 \;\; VS \;\; H_a: \gamma_{frsthalf} \neq 0$$

$$Calculating\ t\ statistics:\ \ t = \frac{-.6609955}{.1081735} = -6.11$$

Rejection Rule: reject H_0 in favor of H_a when the absolute value of t is greater than c (critical value) and c the critical value for the t distribution with 3993=4000-7 df and a 1% significance level. After consulting a t table, we find c=2.576

Decision: since absolute value of t is more than critical value, we have enough evidence to reject the null hypothesis in favour of the alternative. We have sufficient evidence that frsthalf and educ is strongly correlated at 1% significant level. Therefore, we find the relevance condition is satisfied.

(18)

$$\widehat{children} = -3.282776 - .1781478 * educ + .3255732\, age - .0027513 age^2$$

(.6738068) (.0705453) (.0225151) (.0003674)

$$-.11594196 urban - .0435147 electric + .1095386 tv$$

(.0754572) (.1536157) (.2076974)

$n = 4000, R^2 = 0.5467$

There is a quite large differences between whether or not use frsthalf as IV for educ. IV estimates indicates a larger effect of educ on number of children. This indicates that these three proxy variables may not be only source of endogeneity. There might be other sources, including simultaneity and measurement error. The standard error in our IV is much larger than OLS, this is the cost of IV for dealing with endogeneity. So, only when there exists endogenous problem, then we can use IV to get better prediction than OLS. For IV estimate, on average, ceteris paribus, an increase in one year of education gained by women, it is predicted that there is a decrease in 0.178 number of children they have.

(19) Summary: How would education affect women fertility

We use 4000 observations to estimate the effect of education on women fertility.

Firstly, we use the Equation 1 to estimate the effect of educ on children. We find on average, ceteris paribus, an increase in one year of education is associated with an increase in $-.0893658$ number of children. This indicate a negative relationship between year of educ and number of children. By t-test, our results are statistically significant at 1% significant level. However, we doubt that this is a casual effect since there are so many variables that are not included in our estimation, which may lead to omitted variable bias. It may violates the zero conditional mean independence, therefore we cannot conclude a casual effect. We suspect that we omit income which is correlated to fertility and education. By using the correlation, there is a downward bias on estimated coefficient on educ. However, the dataset or data collected do not include family income. We use three proxy variables, including urban, electric and tv, to re-estimate the regression line, getting Eq 2. By looking at the equation 2, we find that the effect of education on fertility becomes smaller, this fits our expectation, proxy variables eliminate the downward bias. Also, we find three proxy variables are jointly significant. Overall, this indicates that these three proxy variables are good proxy. Then, we worry about the heteroscedasticity of our model, we use BP test find that we have sufficient evidence that there exists a heteroscedasticity problem. Therefore, we then use robust se to ensure our inference to be valid.

After that, we still worry about the endogeneity problem, so we use IV method to get consistent estimator. We find on average, if women born in the first six months of the year, the estimated educ=5.45. If not, educ=6.37. We firstly use the reduced form regression to find the Frsthalf is a very relevant IV to education. So, the relevance condition is satisfied. So, we can use IV method. By using it, it has been find a huge difference with the proxy regression we estimated previously. So, apart from omitted variable bias, there exists other sources leading to endogenous. So, using IV is better than the OLS. By looking at the IV regression, the coefficient on educ is statistically significant at 5% level. There is very strong evidence that there is a casual effect of education on fertility. Since after we use proxy and IV, we still find statistically significance. And the effect is tend to be negative.

Appendix:

Untitled.do*

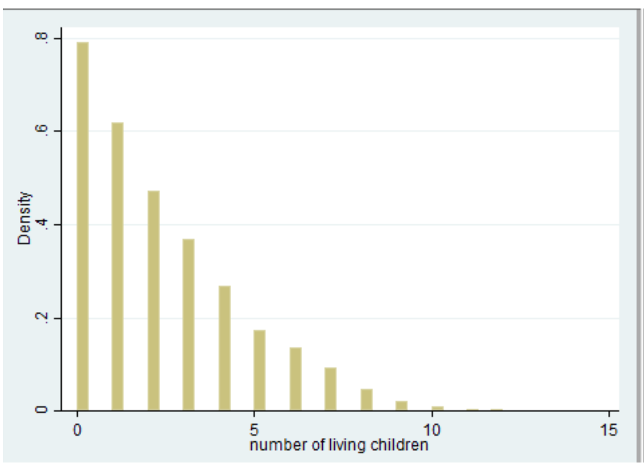
```
summarize children,detail
histogram children
summarize educ,detail
count if educ<1
summarize age agefbrth urban tv electric frsthalf
gen age2=age^2
reg children educ age age2
reg children educ age age2 urban electric tv
predict uhat,residuals
gen uhat2=uhat^2
reg uhat2 educ age age2 urban electric tv
reg children educ age age2 urban electric tv,robust
sum educ if frsthalf==1
sum educ if frsthalf==0
reg educ frsthalf age age2 urban electric tv
ivreg children (educ=frsthalf) age age2 urban electric tv,robust
```

```
. summarize children,detail

                  number of living children
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs               4,000
25%            0              0       Sum of Wgt.       4,000

50%            2                      Mean            2.25975
                        Largest       Std. Dev.      2.218673
75%            4             11
90%            6             12       Variance       4.922511
95%            7             12       Skewness       1.062055
99%            9             12       Kurtosis       3.676533

.
```



. summarize educ,detail

years of education

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	4,000
25%	3	0	Sum of Wgt.	4,000
50%	7		Mean	5.8735
		Largest	Std. Dev.	3.908173
75%	8	19		
90%	10	19	Variance	15.27382
95%	12	19	Skewness	-.0416708
99%	15	20	Kurtosis	2.49733

.
end of do-file

. count if educ<1
815

. summarize age agefbrth urban tv electric frsthalf

Variable	Obs	Mean	Std. Dev.	Min	Max
age	4,000	27.39075	8.712321	15	49
agefbrth	2,989	19.0087	3.110406	10	38
urban	4,000	.51675	.4997818	0	1
tv	4,000	.09275	.290118	0	1
electric	4,000	.13975	.3467708	0	1
frsthalf	4,000	.53975	.4984797	0	1

.

. reg children educ age age2

Source	SS	df	MS
Model	11179.0486	3	3726.34953
Residual	8506.07115	3,996	2.12864643
Total	19685.1198	3,999	4.92251057

Number of obs	=	4,000
F(3, 3996)	=	1750.57
Prob > F	=	0.0000
R-squared	=	0.5679
Adj R-squared	=	0.5676
Root MSE	=	1.459

children	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.0893658	.00623	-14.34	0.000	-.1015801	-.0771515
age	.3321312	.0171985	19.31	0.000	.2984126	.3658498
age2	-.0026402	.0002832	-9.32	0.000	-.0031954	-.002085
_cons	-4.131478	.2500609	-16.52	0.000	-4.621737	-3.641219

.

```
. reg children educ age age2 urban electric tv

      Source |       SS           df       MS      Number of obs   =     4,000
-------------+----------------------------------   F(6, 3993)      =    895.04
       Model |  11290.3266         6  1881.72109   Prob > F        =    0.0000
    Residual |  8394.79318     3,993  2.10237746   R-squared       =    0.5735
-------------+----------------------------------   Adj R-squared   =    0.5729
       Total |  19685.1198     3,999  4.92251057   Root MSE        =      1.45

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0716525   .0067145   -10.67   0.000    -.0848166   -.0584884
         age |   .3398357   .0171272    19.84   0.000     .3062568    .3734145
        age2 |  -.0027289   .0002819    -9.68   0.000    -.0032816   -.0021762
       urban |   -.203995   .0485593    -4.20   0.000    -.2991982   -.1087918
    electric |    -.25055   .0804478    -3.11   0.002    -.4082726   -.0928274
          tv |   -.171123   .0954223    -1.79   0.073     -.358204    .0159581
       _cons |  -4.217031   .2495043   -16.90   0.000    -4.706199   -3.727864
------------------------------------------------------------------------------
. reg uhat2 educ age age2 urban electric tv

      Source |       SS           df       MS      Number of obs   =     4,000
-------------+----------------------------------   F(6, 3993)      =    221.06
       Model |  17124.9546         6  2854.15911   Prob > F        =    0.0000
    Residual |  51554.6904     3,993  12.9112673   R-squared       =    0.2493
-------------+----------------------------------   Adj R-squared   =    0.2482
       Total |  68679.6451     3,999  17.1742048   Root MSE        =    3.5932

------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.1493257   .0166395    -8.97   0.000    -.1819484    -.116703
         age |  -.0751198   .0424439    -1.77   0.077    -.1583335    .0080938
        age2 |    .004623   .0006986     6.62   0.000     .0032533    .0059926
       urban |  -.0756677   .1203375    -0.63   0.530    -.3115964    .1602611
    electric |   .1862056   .1993624     0.93   0.350     -.204656    .5770672
          tv |  -.2996352   .2364717    -1.27   0.205    -.7632517    .1639813
       _cons |   1.255015   .6183111     2.03   0.042     .0427799     2.46725
------------------------------------------------------------------------------
```

```
 reg children educ age age2 urban electric tv,robust

inear regression                                Number of obs     =      4,000
                                                F(6, 3993)        =     885.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5735
                                                Root MSE          =       1.45

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0716525   .0067003   -10.69   0.000    -.0847888   -.0585162
         age |   .3398357   .0199962    17.00   0.000     .3006319    .3790394
        age2 |  -.0027289   .0003658    -7.46   0.000    -.0034461   -.0020116
       urban |   -.203995   .0475067    -4.29   0.000    -.2971347   -.1108553
    electric |    -.25055   .0790036    -3.17   0.002    -.4054412   -.0956588
          tv |   -.171123   .0878761    -1.95   0.052    -.3434091    .0011632
       _cons |  -4.217031   .2533503   -16.65   0.000    -4.713739   -3.720323
------------------------------------------------------------------------------


nd of do-file

. sum educ if frsthalf==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |      2,159    5.450672    3.903878          0         18

. sum educ if frsthalf==0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        educ |      1,841    6.369364    3.855602          0         20

.
end of do-file
```

```
. reg educ frsthalf age age2 urban electric tv

      Source |       SS           df       MS      Number of obs   =     4,000
-------------+----------------------------------   F(6, 3993)      =    214.33
       Model |  14879.5476         6  2479.9246    Prob > F        =    0.0000
    Residual |  46200.4434     3,993  11.570359    R-squared       =    0.2436
-------------+----------------------------------   Adj R-squared   =    0.2425
       Total |   61079.991     3,999  15.2738162   Root MSE        =    3.4015

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    frsthalf |  -.6609955   .1081735    -6.11   0.000     -.873076   -.4489151
         age |  -.1322433    .040125    -3.30   0.001    -.2109107    -.053576
        age2 |  -.0002166   .0006613    -0.33   0.743    -.0015131      .00108
       urban |   .8143062    .113189     7.19   0.000     .5923926     1.03622
    electric |   1.911368   .1863022    10.26   0.000     1.546111    2.276624
          tv |   2.606173    .220025    11.84   0.000     2.174801    3.037544
       _cons |   9.101798   .5713215    15.93   0.000     7.981689    10.22191
------------------------------------------------------------------------------
```

```
 ivreg children (educ=frsthalf) age age2 urban electric tv,robust

Instrumental variables (2SLS) regression          Number of obs   =      4,000
                                                  F(6, 3993)      =     836.24
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.5467
                                                  Root MSE        =     1.4949

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.1781478   .0705453    -2.53   0.012     -.316456   -.0398396
         age |   .3255732   .0225151    14.46   0.000     .2814312    .3697153
        age2 |  -.0027513   .0003674    -7.49   0.000    -.0034716    -.002031
       urban |  -.1159419   .0754572    -1.54   0.124    -.2638802    .0319964
    electric |  -.0435147   .1536157    -0.28   0.777    -.3446872    .2576578
          tv |   .1095386   .2076974     0.53   0.598    -.2976643    .5167415
       _cons |  -3.282776   .6738068    -4.87   0.000    -4.603813   -1.961738
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   age age2 urban electric tv frsthalf
------------------------------------------------------------------------------
```