Estimating Compulsory Schooling Impacts on Labour Market Outcomes in Mexico
Fuzzy Regression Discontinuity Design (RDD) with parametric and non-parametric analyses

Erendira Leon
University of Westminster

2022 UK Stata Conference

Outline

- Applied economics
- Fuzzy RDD
- RDD validity
- Non-parametric analysis
- Parametric analysis
- Conclusions

Applied economics

Analysis of educational policies on **earnings**

- Long debate whether schooling is linked to long-run labour market outcomes
- Measuring the sole impact of education is challenging
- **Endogeneity** between schooling and labour market outcomes: education and earnings are jointly determined
- **Imperfect compliance** with the policy: some factors could affect the exposure to the policy
  a. people not treated that should be treated
  b. people should not be treated and are actually treated

Robust methodology for measuring impact evaluation or the effectiveness of different policies

Fuzzy Regression Discontinuity Design (RDD)

Fuzzy RDD in spirit of Grenet (2013) and Aydemir and Kirdar (2017)

- Non-parametric analysis
- Parametric analysis

Shed light of the **impacts of the 1993 compulsory schooling** on labour market outcomes in Mexico: earnings and employment sectoral choices

- Raise **compulsory school-leaving age from 12 to 15 years**
- Encourage children to **accumulate human capital**

The fuzziness addresses **imperfect compliance** with the policy

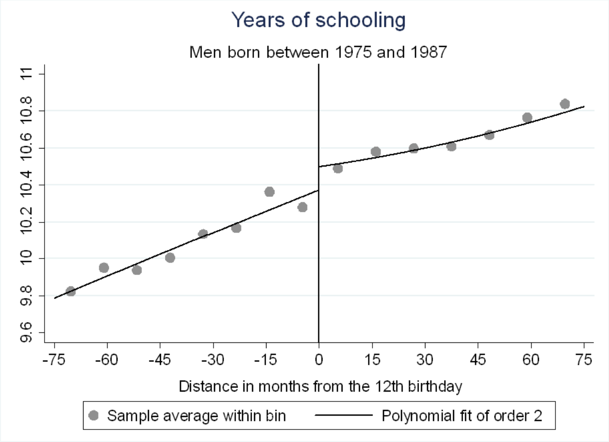
- Use the random assignment of the exposure to the policy

Fuzzy Regression Discontinuity Design (RDD)

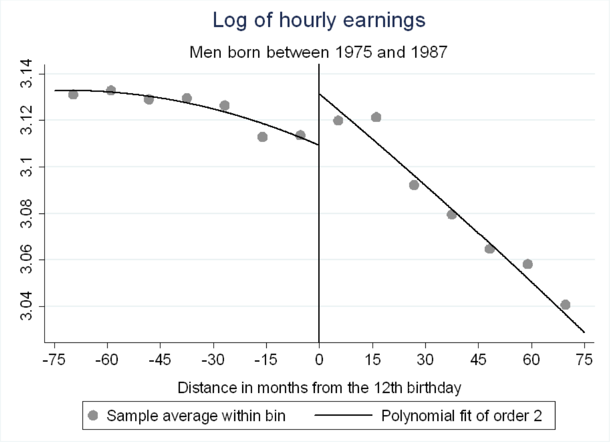
- Age cohort discontinuities measured in **months of birth**
- **Exogenous extra-compulsory schooling** faced by different birth cohorts
- Compare people **treated with untreated** by the policy
- **Running variable** is the age in months of birth from the cohort born in September 1981

$$Treatment_i \begin{cases} 1, & if\ cohort\ born \geq September\ 1981 \\ 0, & if\ cohort\ born < September\ 1981 \end{cases}$$

RDD validity -Discontinuity plots



RDD validity -Discontinuity plots



RDD validity -Discontinuity plots

rdplot implements several data-driven regression-discontinuity (RD) plots, using either evenly spaced or quantile-spaced partitioning

> **rdplot** *depvar* *runvar* [*if*] [*in*] [, c(*cutoff*) p(*pvalue*) binselect(*binmethod*) graph_options(*gphopts*)]

where *depvar* is the dependent variable, and *runvar* is the running variable (also known as the score or forcing variable).

c(*cutoff*) specifies the RD cutoff. The default is c(0).

RDD validity -Discontinuity plots

rdplot implements several data-driven regression-discontinuity (RD) plots, using either evenly spaced or quantile-spaced partitioning

rdplot *depvar* *runvar* [*if*] [*in*] [, c(*cutoff*) p(*pvalue*) binselect(*binmethod*) graph_options(*gphopts*)]

where *depvar* is the dependent variable, and *runvar* is the running variable (also known as the score or forcing variable).

c(*cutoff*) specifies the RD cutoff. The default is c(0).

p(*pvalue*) for the order of the global polynomial used to approximate the population conditional mean functions. The default is p(4).

RDD validity -Discontinuity plots

rdplot implements several data-driven regression-discontinuity (RD) plots, using either evenly spaced or quantile-spaced partitioning

rdplot *depvar* *runvar* [*if*] [*in*] [, c(*cutoff*) p(*pvalue*) binselect(*binmethod*) graph_options(*gphopts*)]

where *depvar* is the dependent variable, and *runvar* is the running variable (also known as the score or forcing variable).

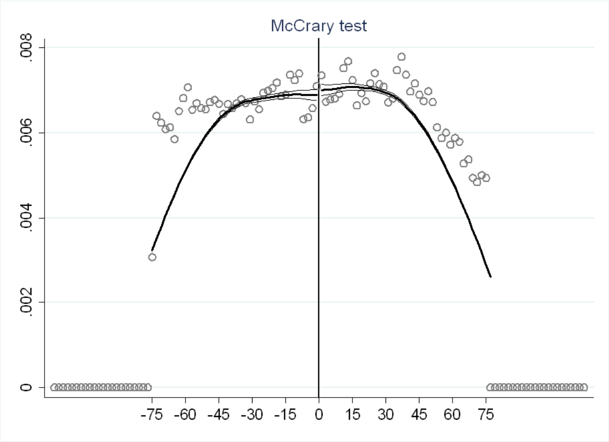
c(*cutoff*) specifies the RD cutoff. The default is c(0).

p(*pvalue*) for the order of the global polynomial used to approximate the population conditional mean functions. The default is p(4).

binselect(*binmethod*) for selecting the number of bins. E.g., **es** specifies the optimal evenly spaced method using spacings estimators.

graph_options(*gphopts*) graphical options

RDD validity -McCrary test



RDD validity -McCrary test

DCdensity implements standard sufficient conditions for identification in the regression discontinuity design continuity of the conditional expectation of counterfactual outcomes in the running variable.

> **DCdensity** *Z*, breakpoint(0) generate(Xj Yj r0 fhat se_fhat) graphname(DCdensity_example.eps)

where *Z* is the running variable

breakpoint for the threshold/cutoff value in the running var, which determines the two samples (e.g., control and treatment units in RD settings). The default is (0)

local linear smoother on the scatterplot (Xj, Yj), *r0* for the values above and below the running var, *fhat* estimation of the density function, and *se_fhat* the standard errors of the estimation of the density function

Stata in applied economics: Fuzzy RDD

Fuzzy Regression Discontinuity Design (RDD)

First stage

$$Years\ of\ Schooling_i = \alpha_0 + \alpha_1(Treatment_i) + \alpha_2 F(Age\ in\ months_i) + \alpha_3 X_i + \varepsilon_i \quad (1)$$

Reduced-form

$$LMkt\ outcomes_i = \beta_0 + \beta_1(Treatment_i) + \beta_2 F(Age\ in\ months_i) + \beta_3 X_i + \omega_i \quad (2)$$

Second stage: 2SLS

$$LMkt\ outcomes_i = \delta_0 + \delta_1\left(\widehat{Years\ of\ Schooling_i}\right) + \delta_2 F(Age\ in\ months_i) + \delta_3 X_i + \mu_i \quad (3)$$

X_i survey year dummies, birth states dummies, urban status, economic sector

Non-parametric analysis: rdbwselect and rdrobust

rdbwselect implements bandwidth selectors for local-polynomial RD estimators proposed in Calonico, Cattaneo, and Titiunik (2014). It also computes the bandwidth selection procedures

rdbwselect *depvar runvar* [*if*] [*in*] [,c(*cutoff*) p(*pvalue*) q(*qvalue*) rho(*rhovalue*) kernel(*kernelfn*) bwselect(*bwmethod*) vce(*vcemethod*) all]

Non-parametric analysis: rdbwselect and rdrobust

rdbwselect implements bandwidth selectors for local-polynomial RD estimators proposed in Calonico, Cattaneo, and Titiunik (2014). It also computes the bandwidth selection procedures

> **rdbwselect** *depvar runvar* [*if*] [*in*] [,c(*cutoff*) p(*pvalue*) q(*qvalue*) rho(*rhovalue*) kernel(*kernelfn*) bwselect(*bwmethod*) vce(*vcemethod*) all]

rdrobust implements local-polynomial RD point estimators with robust confidence intervals proposed in Calonico, Cattaneo, and Titiunik (2014)

> **rdrobust** *depvar runvar* [*if*] [*in*] [,c(*cutoff*) p(*pvalue*) q(*qvalue*) **fuzzy(fuzzyvar)** kernel(*kernelfn*) h(*hvalue*) b(*bvalue*) rho(*rhovalue*) bwselect(*bwmethod*) delta(*deltavalue*) vce(*vcemethod*) level(*level*) all]

Non-parametric analysis: rdbwselect and rdrobust

q(*qvalue*) for the order of the local polynomial used to construct the bias correction. The default is q(2) (local quadratic regression).

rho(*rhovalue*) sets the pilot bandwidth, b_n, equal to h_n/rho, where h_n is computed using the method and options chosen below.

kernel(*kernelfn*) specifies the kernel function used to construct the local polynomial estimators. Options are triangular, epanechnikov, and uniform. The default is kernel(triangular)

fuzzy(*fuzzyvar*) for the treatment status variable implementing **fuzzy RD estimation**. The default is sharp RD design. For fuzzy RD designs, bandwidths are estimated using sharp RD bandwidth selectors for the reduced-form outcome equation.

Non-parametric analysis: Results

The evidence suggests that although **the policy raises years of schooling it did not exert impacts on labour market earnings**

Estimation method	First-stage				Reduced-form				2SLS			
Dependent variable	Years of schooling				Log of hourly earnings				Log of hourly earnings			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Treatment	0.288**	0.277*	0.275**	0.236*	0.024	0.024	0.016	0.015				
	(0.142)	(0.145)	(0.125)	(0.132)	(0.020)	(0.021)	(0.018)	(0.019)				
$\widehat{\text{Years of schooling}}$									0.086	0.085	0.060	0.062
									(0.068)	(0.073)	(0.063)	(0.080)
Obs.	145,035	145,035	145,035	145,035	145,035	145,035	145,035	145,035	145,035	145,035	145,035	145,035
Eff. Number of obs.	37,447	35,442	47,611	39,454	37,447	35,442	47,611	39,454	37,447	35,442	47,611	39,454
Optimal bandwidth	32.13	31.25	38.64	33.90	32.13	31.25	38.64	33.90	32.13	31.25	38.64	33.90
Survey year dummies	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Birth region dummies	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Urban status	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes

Notes: *p<0.1, ** p<0.05, *** p<0.01

The sample is constructed from the 2009-2017 Mexican National Occupations and Employment Survey. Following Calonico et al. (2018) and Calonico et al. (2014) for the optimal bandwidth. Robust standard errors using EHW correction as recommended by Kolesár and Rothe (2018) in parentheses.

Parametric analysis: 2SLS, reg, iveg2

Similar to a Two-Stage Least-Squares regression (2SLS)

- **First stage**

regress performs ordinary least-squares linear regression. It can also compute robust and cluster-robust standard errors.

regress *depvar* [*indepvars*] [*if*] [*in*] [weight] [, *options*]

where *depvar* is the dependent variable, the exogenous variable or instrument: *years of schooling*

indepvars are independent variables: the running variable, and interacted quadratic specifications for the running variable with the treatment variable on both sides of the threshold

options for the type of standard error reported. E.g., *robust, cluster*, etc.

Parametric analysis: 2SLS, reg, iveg2

- **Reduced-form**

Similar...

regress *depvar* [*indepvars*] [*if*] [*in*] [weight] [, *options*]

- **IV 2SLS**

ivreg2 implements a range of single-equation estimation methods for the linear regression model: ordinary least squares (OLS), instrumental variables (IV, also known as two-stage least squares, 2SLS), the generalized method of moments (GMM), etc

ivreg2 *depvar* [*varlist1*] (*varlist2* = *varlist_iv*) [*if*] [*in*] [*weight*] [,*options*]

Parametric analysis: 2SLS, reg, iveg2

varlist1 are the exogenous regressors or included instruments

varlist_iv are the exogenous variables excluded from the regression or excluded instruments

varlist2 the endogenous regressors that are being instrumented, the treatment group

Parametric analysis: Results

There is no empirical evidence to suggest that the policy exerts impacts on labour market earnings

Interacted quadratic specification

Estimation method	First-stage				Reduced-form				2SLS			
Dependent variable	Years of schooling				Log of hourly wages				Log of hourly wages			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Treatment	0.147*	0.147*	0.137*	0.116	0.016	0.016	0.015	0.012				
	(0.082)	(0.082)	(0.081)	(0.079)	(0.012)	(0.012)	(0.011)	(0.011)				
$\widehat{\text{Years of schooling}}$									0.110	0.109	0.110	0.106
									(0.075)	(0.075)	(0.080)	(0.094)
Obs.	85,890	85,890	85,890	85,890	85,890	85,890	85,890	85,890	85,890	85,890	85,890	85,890
Survey year dummies	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Birth region dummies	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Urban status	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes

*p<0.1, ** p<0.05, *** p<0.01

Robust standard errors correction as recommended by Kolesár and Rothe (2018)

Conclusions

- Fuzzy RDD implemented with Stata **to analyse policy impacts**
- Different tests can be applied with Stata for **validating** the implementation of Fuzzy RDD
  - RDD plots (rdplot)
  - Mccrary test (DCdensity)
- Stata allows the **non-parametric and parametric analysis**
  - rdrobust
  - rdbwselect
  - ivreg2

Thank you!

Reference

Aydemir, A., Kirdar, M. G. (2017), "Low Wage Returns to Schooling in a Developing Country: Evidence from a Major Policy Reform in Turkey", Oxford Bulletin of Economics and Statistics, 79(6), 1046–1086.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014), "Robust nonparametric confidence intervals for regression-discontinuity designs", Econometrica.

Grenet, J. (2013), "Is Extending Compulsory Schooling Alone Enough to Raise Earnings? Evidence from French and British Compulsory Schooling Laws", Scandinavian Journal of Economics, 115(1), 176–210.

McCrary, J (2008), "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test", Journal of Econometrics

https://eml.berkeley.edu/~jmccrary/mccrary2006_DCdensity.pdf
https://eml.berkeley.edu/~jmccrary/DCdensity/

Data

National Employment Survey (ENOE) from 2009 to 2017

- Report, *inter alia*, age in months, years of schooling, earnings, etc
- Male observations aged between 24 to 40 years when surveyed
- Born between 1975 and 1987 and aged in a range of 6-18 years at the time of the reform

Example: Non-parametric Stata commands

```
foreach var of varlist lg_inc {
  2. rdbwselect `var' arecen if $sample2b, fuzzy(year_sch) kernel(tri) all
vce(hc2) bwselect(mserd)
  3. global `var'_bw1 = e(b_mserd)
  4. global `var'_bw2 = e(h_mserd)
  5.
. forvalues z=1(1)1 {
  6. local n= `z' + 1
  7.
. rdrobust `var' arecen if $sample2b, fuzzy(year_sch) kernel(tri) all
vce(hc2) bwselect(mserd) h(${`var'_bw`n'}) b(${`var'_bw`z'}) p(2)
  8. test Conventional
  9. test Bias
 10. test Robust
 11.
 12. }
```

Example: Non-parametric Stata output

```
Bandwidth estimators for fuzzy RD local polynomial regression.

     Cutoff c = 0 | Left of c  Right of c            Number of obs =     148964
------------------+----------------------            Kernel        = Triangular
    Number of obs |     74618       74346            VCE method    =        HC2
    Min of arecen |   -75.000       0.000
    Max of arecen |    -1.000      75.000
   Order est. (p) |         1           1
   Order bias (q) |         2           2

Outcome: lg_inc. Running variable: arecen. Treatment Status: year_sch.
--------------------------------------------------------------------------------
                  |       BW est. (h)          |        BW bias (b)
           Method | Left of c       Right of c | Left of c       Right of c
------------------+----------------------------+--------------------------------
            mserd |    25.747          25.747  |    44.446          44.446
           msetwo |    16.950          28.188  |    31.721          38.319
           msesum |    20.930          20.930  |    35.719          35.719
         msecomb1 |    20.930          20.930  |    35.719          35.719
         msecomb2 |    20.930          25.747  |    35.719          38.319
------------------+----------------------------+--------------------------------
            cerrd |    14.193          14.193  |    44.446          44.446
           certwo |     9.344          15.539  |    31.721          38.319
           cersum |    11.538          11.538  |    35.719          35.719
         cercomb1 |    11.538          11.538  |    35.719          35.719
         cercomb2 |    11.538          14.193  |    35.719          38.319
--------------------------------------------------------------------------------
```

Example: Non-parametric Stata output

```
Fuzzy RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =     148964
-------------------+----------------------            BW type       =     Manual
     Number of obs |     74618       74346            Kernel        = Triangular
Eff. Number of obs |     25876       27383            VCE method    =        HC2
    Order est. (p) |         2           2
    Order bias (q) |         3           3
       BW est. (h) |    25.747      25.747
       BW bias (b) |    44.446      44.446
         rho (h/b) |     0.579       0.579

First-stage estimates. Outcome: year_sch. Running variable: arecen.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|     [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  .24941     .11274   2.2124   0.027    .028454      .470372
    Bias-corrected |  .26205     .11274   2.3245   0.020    .041094      .483012
            Robust |  .26205     .12038   2.1769   0.029     .02611      .497996
--------------------------------------------------------------------------------

Treatment effect estimates. Outcome: lg_inc. Running variable: arecen. Treatment Status: year_sch.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|     [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  .06596     .06214   1.0615   0.288   -.055834      .187763
    Bias-corrected |  .05903     .06214   0.9498   0.342   -.062773      .180824
            Robust |  .05903     .06641   0.8888   0.374   -.071138      .189189
--------------------------------------------------------------------------------
```

Example: Non-parametric Stata output

```
Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =     148964
-------------------+----------------------            BW type       =     Manual
     Number of obs |     74618       74346            Kernel        = Triangular
Eff. Number of obs |     25876       27383            VCE method    =        HC2
    Order est. (p) |         2           2
    Order bias (q) |         3           3
       BW est. (h) |    25.747      25.747
       BW bias (b) |    44.446      44.446
         rho (h/b) |     0.579       0.579

Outcome: lg_inc. Running variable: arecen.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  .01645     .01721   0.9558   0.339   -.017284      .050188
    Bias-corrected |  .01556     .01721   0.9037   0.366   -.018181      .049292
            Robust |  .01556     .01839   0.8458   0.398   -.020492      .051603
--------------------------------------------------------------------------------
```

Example: Parametric Stata commands

```
*First stage
*Spline - Quadratic specification
reg year_sch aTER arecenaTER arecen2aTER arecenaTER_UT arecen2aTER_UT,
robust

*Reduced form
*Spline - Quadratic specification
reg lg_inc aTER arecenaTER arecen2aTER arecenaTER_UT arecen2aTER_UT, robust

*Second stage
*Spline Quadratic specification
ivreg2 lg_inc (year_sch = aTER) arecenaTER arecen2aTER arecenaTER_UT
arecen2aTER_UT, robust endog (year_sch)
```

Example: Parametric Stata output

First stage

```
Linear regression                               Number of obs     =     82,125
                                                F(5, 82119)       =      37.97
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0023
                                                Root MSE          =     4.0209

------------------------------------------------------------------------------
             |               Robust
    year_sch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aTER |   .1658821   .0854494     1.94   0.052    -.0015982    .3333624
   arecenaTER |   .0033887   .0065208     0.52   0.603    -.0093921    .0161695
  arecen2aTER |   .0000339   .0001599     0.21   0.832    -.0002795    .0003473
arecenaTER_UT |  -.0006796   .0074534    -0.09   0.927    -.0152881     .013929
arecen2aTER_UT |  -.0002252   .0001806    -1.25   0.212    -.0005793    .0001288
       _cons |    10.3233   .0648346   159.23   0.000     10.19622    10.45037
------------------------------------------------------------------------------
```

Example: Parametric Stata output

Reduced-form

```
Linear regression                               Number of obs     =     82,125
                                                F(5, 82119)       =       9.21
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0005
                                                Root MSE          =     .61498

------------------------------------------------------------------------------
             |               Robust
      lg_inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aTER |   .0170504    .013021     1.31   0.190    -.0084706    .0425714
   arecenaTER |  -.0007443   .0009899    -0.75   0.452    -.0026846    .0011959
  arecen2aTER |  -.0000159   .0000244    -0.65   0.514    -.0000636    .0000318
arecenaTER_UT |   -.000218    .001136    -0.19   0.848    -.0024446    .0020086
arecen2aTER_UT |  8.67e-06   .0000276     0.31   0.754    -.0000455    .0000628
       _cons |   3.111498   .0098806   314.91   0.000     3.092132    3.130864
------------------------------------------------------------------------------
```

Example: Parametric Stata output

```
IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                      Number of obs =    82125
                                                      F(  5, 82119) =    10.35
                                                      Prob > F      =   0.0000
Total (centered) SS     =  31073.90264                Centered R2   =   0.1278
Total (uncentered) SS   =  826807.4245                Uncentered R2 =   0.9672
Residual SS             =  27102.96524                Root MSE      =    .5745

------------------------------------------------------------------------------
             |               Robust
      lg_inc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    year_sch |   .1027862   .0731135     1.41   0.160    -.0405136     .246086
  arecenaTER |  -.0010927   .0010853    -1.01   0.314    -.0032198    .0010345
 arecen2aTER |  -.0000194   .0000219    -0.89   0.375    -.0000623    .0000235
arecenaTER_UT|  -.0001482   .0010205    -0.15   0.885    -.0021484     .001852
arecen2aTER_UT|  .0000318   .0000209     1.52   0.128    -9.13e-06    .0000728
       _cons |   2.050406   .7617748     2.69   0.007     .5573543    3.543457
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):              3.768
                                                   Chi-sq(1) P-val =    0.0523
------------------------------------------------------------------------------
Weak identification test (Kleibergen-Paap rk Wald F statistic):          3.769
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.000
                                                 (equation exactly identified)
-endog- option:
Endogeneity test of endogenous regressors:                               0.273
                                                   Chi-sq(1) P-val =    0.6015
Regressors tested:    year_sch
------------------------------------------------------------------------------
Instrumented:         year_sch
Included instruments: arecenaTER arecen2aTER arecenaTER_UT arecen2aTER_UT
Excluded instruments: aTER
------------------------------------------------------------------------------
```