DEMO: Regression discontinuity analysis

```
/* Workshop of program impact evaluation *
 ****************************************************************************************

. *****************

. * This example deals with a subset of the PROGRESA/Oportunidades dataset. Only women 20 to
24 years of age are represented here. The hypothesis we were trying to test was whether the
program had an effect of increasing the use of contraception in young adult women.
.
. * However, that was not the only reason to conduct this analysis: since PROGRESA was an
experiment, we were interested in comparing different methods against the "gold standard",
particularly Regression discontinuity. Remember that RDA is useful when participation on the
program is defined by a threshold in a continuous variable, in this case such variable is the
"poverty score" and the threshold was 752 points.

    * The data here were gathered in the year 2000, after 3 years of the start of the
    experiment.

***** The first part is to get to know the data, particularly, lets obtain summary statistics
of our outcome (current use of contraception), how many are assigned to a program  area, and
how many of those eligible.
.
. desc

Contains data from C:\

  obs:         2,239
 vars:            17                          14 Apr 2011 09:50
 size:       105,233 (99.0% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
folio           long   %12.0g                 household ID
line            byte   %8.0g                  person ID
locality        float  %9.0g                  locality ID
state           byte   %8.0g       entidad    state (province)
program         byte   %8.0g       yn         program or control locality
score           float  %9.0g                  poverty score (1997)
eligible        byte   %8.0g       elig       eligible to participation
age             byte   %8.0g       p08        age in years
any_schooling   byte   %8.0g       yesno      do subject have any schooling?
goes_school     byte   %8.0g       yesno      currently goes to school?
dich_job        float  %9.0g       yesno      does subject have a job?
income          float  %9.0g                  monthly income in 2006 USD
wave            float  %9.0g                  yearndata was collected
num_child       byte   %8.0g                  total live born children
pregnant        byte   %8.0g       yesno      is subject currenltly pregnant?
any_contracep   byte   %8.0g       yesno      ever used contraception?
dich_contracep  float  %9.0g       yesno      currently uses contraception?
-------------------------------------------------------------------------------
Sorted by:  folio  line  wave
     Note:  dataset has changed since last saved

.
. tab program eligible, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

program or |      eligible to
   control |    participation
 locality  | non-eligi   eligible |     Total
-----------+----------------------+----------
   control |       429        449 |       878
           |     48.86      51.14 |    100.00
```

```
-----------+----------------------+----------
   program |       657        704 |     1,361
           |     48.27      51.73 |    100.00
-----------+----------------------+----------
     Total |     1,086      1,153 |     2,239
           |     48.50      51.50 |    100.00


. summ score

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       score |      2239    755.8282    132.6151        274       1246

. kdensity score, xline(752)

.
. ***** Now let's obtain our estimate of the program effect in the "usual" way, given that
this is an experiment, let's fit a OLS model with current use of contraception as the outcome
variable. Covariates to adjust for will be age, state (province), and the poverty score
(remember we only are supposed to adjust for baseline covariates or covariates that dont
change in time). Remember we need to take care of intra-cluster correlation and
heteroskedasticity by calculating robust standard errors clustering around locality. Also
remember that we want to focus only on the eligibles.

. xi: reg dich_contracep program i.state score age if eligible==1, cluster(locality)
i.state           _Istate_12-30       (naturally coded; _Istate_12 omitted)

Linear regression                                      Number of obs =    1153
                                                       F(  8,   339) =   13.25
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0722
                                                       Root MSE      =  .39616

                              (Std. Err. adjusted for 340 clusters in locality)
------------------------------------------------------------------------------
             |               Robust
dich_contr~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     program |   .0677429   .0268607     2.52   0.012     .0149083    .1205775
  _Istate_13 |   .2236205   .0457285     4.89   0.000      .133673    .3135679
  _Istate_21 |   .1580735    .034607     4.57   0.000      .090002    .2261451
  _Istate_22 |     .20816   .0656321     3.17   0.002     .0790627    .3372574
  _Istate_24 |    .297262   .0556132     5.35   0.000     .1878716    .4066524
  _Istate_30 |   .1825666   .0268034     6.81   0.000     .1298447    .2352884
       score |  -.0001719   .0001975    -0.87   0.385    -.0005604    .0002167
         age |   .0384747   .0082449     4.67   0.000      .022257    .0546924
       _cons |  -.7442159   .2443301    -3.05   0.003     -1.22481   -.2636219
------------------------------------------------------------------------------

.
. * What is the estimate of the effect of the program? Interpret it
.
. * This estimate is our gold standard. Now let’s run some RDA models and see how they go,
. * but before lets define windows of 25, 50 and 75 points around the threshold
.
.         gen w25=abs(score-752)<=25

.         gen w50=abs(score-752)<=50

.         gen w75=abs(score-752)<=75

.
. * Lets run the regressions now (notice that now we don’t include the control group: we are
pretending not to have it!, also notice that "eligible" takes the place of "program" in the
model.


. xi: reg dich_contracep eligible i.state score age if program==1 & w25==1, cluster(locality)
i.state           _Istate_12-30       (naturally coded; _Istate_12 omitted)

Linear regression                                      Number of obs =     280
```

```
                                                F(  8,   135) =    4.42
                                                Prob > F      =  0.0001
                                                R-squared     =  0.0765
                                                Root MSE      =  .39836

                           (Std. Err. adjusted for 136 clusters in locality)
------------------------------------------------------------------------------
             |               Robust
dich_contr~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    eligible |  -.2255288   .0863617    -2.61   0.010    -.3963257    -.054732
  _Istate_13 |    .234072    .067457     3.47   0.001     .1006629    .3674811
  _Istate_21 |   .1856756   .0771704     2.41   0.017     .0330563    .3382949
  _Istate_22 |   .0794319    .101293     0.78   0.434    -.1208945    .2797584
  _Istate_24 |    .237356   .0771617     3.08   0.003      .084754     .389958
  _Istate_30 |   .1688011   .0553657     3.05   0.003     .0593048    .2782975
       score |  -.0034954   .0030586    -1.14   0.255    -.0095443    .0025535
         age |   .0104456   .0154068     0.68   0.499    -.0200243    .0409154
       _cons |   2.560219   2.352072     1.09   0.278    -2.091455    7.211893
------------------------------------------------------------------------------

.         xi: reg dich_contracep eligible i.state score age if program==1 & w50==1,
cluster(locality)
i.state           _Istate_12-30       (naturally coded; _Istate_12 omitted)

Linear regression                               Number of obs =     489
                                                F(  8,   171) =    8.46
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0738
                                                Root MSE      =   .4056

                           (Std. Err. adjusted for 172 clusters in locality)
------------------------------------------------------------------------------
             |               Robust
dich_contr~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    eligible |  -.1859991   .0681453    -2.73   0.007    -.3205135   -.0514848
  _Istate_13 |    .260209   .0496229     5.24   0.000     .1622567    .3581612
  _Istate_21 |   .1866252   .0588726     3.17   0.002     .0704145    .3028359
  _Istate_22 |   .1677732   .0689834     2.43   0.016     .0316045    .3039419
  _Istate_24 |   .3066243   .0684165     4.48   0.000     .1715746     .441674
  _Istate_30 |   .1377662    .036521     3.77   0.000     .0656761    .2098563
       score |   -.003127   .0013736    -2.28   0.024    -.0058384   -.0004155
         age |   .0220183   .0119312     1.85   0.067     -.001533    .0455697
       _cons |   2.003005   1.114413     1.80   0.074    -.1967726    4.202782
------------------------------------------------------------------------------

.         xi: reg dich_contracep eligible i.state score age if program==1 & w75==1,
cluster(locality)
i.state           _Istate_12-30       (naturally coded; _Istate_12 omitted)

Linear regression                               Number of obs =     655
                                                F(  8,   192) =   12.35
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0698
                                                Root MSE      =  .40807

                           (Std. Err. adjusted for 193 clusters in locality)
------------------------------------------------------------------------------
             |               Robust
dich_contr~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    eligible |  -.0945223   .0613571    -1.54   0.125    -.2155428    .0264982
  _Istate_13 |   .2671395   .0469816     5.69   0.000     .1744732    .3598057
  _Istate_21 |   .1919925   .0456503     4.21   0.000      .101952     .282033
  _Istate_22 |   .1915306   .0826135     2.32   0.021     .0285842    .3544771
  _Istate_24 |    .332922   .0656224     5.07   0.000     .2034886    .4623553
  _Istate_30 |   .1710412   .0333693     5.13   0.000     .1052237    .2368588
       score |  -.0013187   .0008332    -1.58   0.115    -.0029622    .0003247
         age |   .0304219   .0110356     2.76   0.006     .0086553    .0521884
       _cons |   .3984986    .720761     0.55   0.581    -1.023128    1.820125
------------------------------------------------------------------------------
```

```
.
. * What to do you think about these results? Why do they apparently contradict the
experimental design ITT estimate?
```

The problem we see here is **impact heterogeneity.** There is a negative impact near the threshold but a large positive impact in the poorest of the poor. In average the impact is positive. This illustrates one of the main imitations of RDD.