Regression II
Vanderbilt University
Prof. Sean P. Corcoran

December 2, 2023

Lecture 8 In-Class Example *Solutions*

Example 1. This example is taken from *Mastering Metrics* chapter 4 and is based on Carpenter & Dobkin (2009). The analysis revisits the question of whether legal access to alcohol is associated with higher mortality rates. The dataset referenced below includes death rates by age, in which age (from 19-23) is divided into 50 equal-width cells. Individuals are considered "treated" when they reach 21 and can legally drink alcohol. The discontinuity in treatment is sharp, since *all* persons who reach this age are treated. If legal access to alcohol is associated with higher mortality, we would expect to see a discontinuity in death rates at 21. Read the file *AEJfigs.dta* into Stata using:

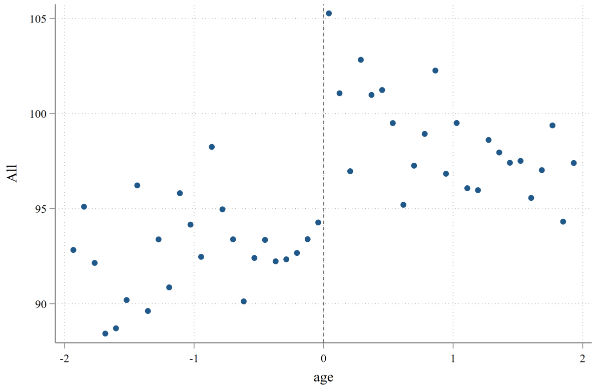
```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/AEJfigs.dta
```

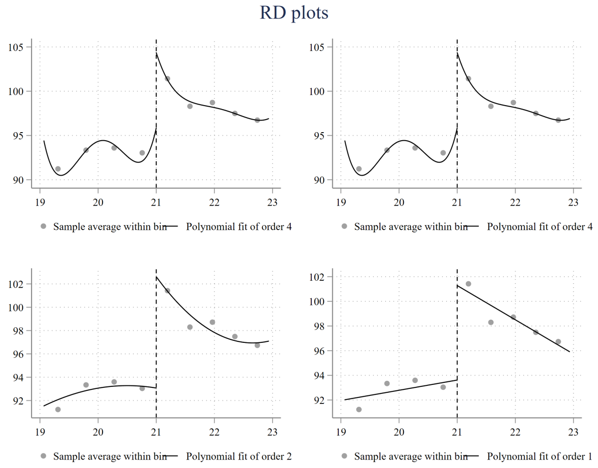
(a) Create a new age variable *age* that is centered at 21, and a "treatment" variable *over21* that equals 1 if *age* $>$ 0 (and 0 otherwise).

```
gen age = agecell - 21
gen over21 = agecell >= 21
```

(b) Create a scatterplot of mortality rates from all cases (*all*) against *age*. Is a discontinuity evident? Try a few RD plots using different bin methods (e.g., ES or QS, IMSE or MV) and different polynomial orders (e.g., quartic (the default), quadratic, linear). Note the RD plot does not have a lot of added value here, since the data are already binned.

The scatterplot and RD plots are shown below. In all five plots a discontinuity is evident at age 21. Again, since the data are already binned, the scatterplot by itself is quite revealing of the discontinuity.



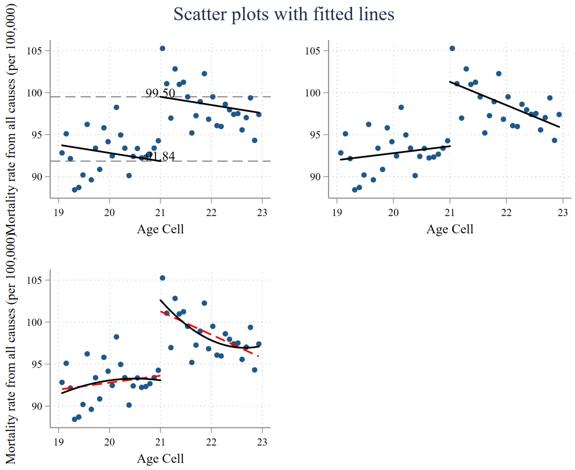

RD plots

(c) We will eventually use `rdrobust` for estimation. But for now use OLS regression to estimate the RD under different scenarios:

- Linear model, assuming the same slope on both sides
- Linear model, assuming different slopes on each side
- Quadratic model, assuming the same slope terms on both sides
- Quadratic model, assuming different slope terms on each side

How are the coefficients interpreted in these models?

Get predicted values and overlay these predicted values on the scatterplot. (Note: this is what `rdplot` can do for you, but this is for illustration). You can easily reproduce Figure 4.2 in *Mastering Metrics* (linear model, same slopes).


Scatter plots with fitted lines

Regression results shown below. In the first model, the slope on age is assumed to be the same on both sides of 21. This is estimated to be -0.975, meaning all-cause mortality is declining with age. The intercept shift at 21 is 7.66, meaning there is a jump in all-causes mortality at age 21 of 7.66. Note that *age* is centered at 21, so the constant term of 91.8 is the predicted all-cause mortality rate when *age*=0 (i.e., age 21)

In the second model, the close on age is allowed to differ on each side of 21. The slope below 21 is estimated to be 0.827; above 21 the slope is -3.60 *lower* (the interaction term). The discontinuity at age 21 is a very similar 7.66.

The third model is a quadratic with the linear and quadratic terms slopes assumed to be the same on each side of 21. In the fourth model, the slopes are allowed to differ on each side of 21. The interaction terms indicate how *different* the slope terms are above 21 vs. below. The discontinuity in the third model is 7.66; in the fourth it is a larger 9.55.

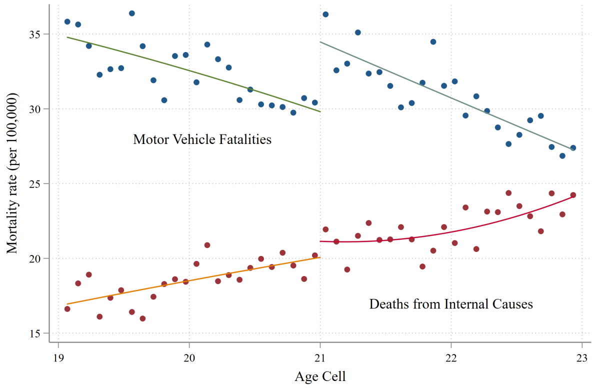
```
.         // linear model in age with intercept shift at over21
.         reg all age over21

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(2, 45)        =     32.99
       Model |  410.138151         2  205.069075   Prob > F        =    0.0000
    Residual |  279.682408        45  6.21516463   R-squared       =    0.5946
-------------+----------------------------------   Adj R-squared   =    0.5765
       Total |  689.820559        47  14.6770332   Root MSE        =     2.493

------------------------------------------------------------------------------
         all |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.9746843   .6324613    -1.54   0.130    -2.248527    .2991581
      over21 |   7.662709   1.440286     5.32   0.000     4.761824    10.56359
       _cons |   91.84137   .8050394   114.08   0.000     90.21994     93.4628
------------------------------------------------------------------------------

.         predict allfitlin
(option xb assumed; fitted values)

.         // linear model in age with intercept shift at over21, different slope
.         // below and above c
.         reg all c.age##i.over21

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(3, 44)        =     29.47
       Model |  460.574058         3  153.524686   Prob > F        =    0.0000
    Residual |  229.246501        44  5.21014775   R-squared       =    0.6677
-------------+----------------------------------   Adj R-squared   =    0.6450
```

```
       Total |  689.820559        47  14.6770332   Root MSE        =    2.2826

------------------------------------------------------------------------------
         all |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .8269952   .8189316     1.01   0.318     -.823453    2.477443
    1.over21 |   7.662709   1.318704     5.81   0.000     5.005035    10.32038
             |
over21#c.age |
          1  |  -3.603359   1.158144    -3.11   0.003    -5.937445   -1.269273
             |
       _cons |   93.61837   .9324647   100.40   0.000     91.73911    95.49763
------------------------------------------------------------------------------

.         predict allfitlini
(option xb assumed; fitted values)

.         reg all c.age##c.age over21

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(3, 44)        =     28.12
       Model |  453.339903         3  151.113301   Prob > F        =    0.0000
    Residual |  236.480656        44  5.37456036   R-squared       =    0.6572
-------------+----------------------------------   Adj R-squared   =    0.6338
       Total |  689.820559        47  14.6770332   Root MSE        =    2.3183

------------------------------------------------------------------------------
         all |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.9746843   .5881378    -1.66   0.105    -2.159998    .2106296
             |
 c.age#c.age |  -.8186505   .2887482    -2.84   0.007    -1.400584   -.2367167
             |
      over21 |   7.662709   1.339349     5.72   0.000     4.963428    10.36199
       _cons |   92.90274   .8370061   110.99   0.000     91.21587    94.58962
------------------------------------------------------------------------------

.         predict allfitq
(option xb assumed; fitted values)

.         // quadratic model in age with intercept shift at over21, different slopes
.         // below and above c
.         reg all c.age##c.age##i.over21

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(5, 42)        =     18.02
       Model |  470.512103         5  94.1024205   Prob > F        =    0.0000
    Residual |  219.308457        42  5.22162992   R-squared       =    0.6821
-------------+----------------------------------   Adj R-squared   =    0.6442
       Total |  689.820559        47  14.6770332   Root MSE        =    2.2851
```

```
------------------------------------------------------------------------------
         all |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.8305827   3.290064    -0.25   0.802    -7.470202    5.809036
             |
 c.age#c.age |  -.8402999   1.615268    -0.52   0.606    -4.100043    2.419443
             |
    1.over21 |   9.547789   1.985277     4.81   0.000     5.541337    13.55424
             |
over21#c.age |
           1 |  -6.017014   4.652854    -1.29   0.203    -15.40685    3.372825
             |
over21#c.age#|
       c.age |
           1 |   2.904189   2.284334     1.27   0.211    -1.705784    7.514162
             |
       _cons |   93.07294   1.403803    66.30   0.000     90.23995    95.90593
------------------------------------------------------------------------------

.         predict allfitqi
```

(d) Repeat the linear and quadratic models for two other measures of mortality: deaths by motor vehicle accidents and deaths by internal causes. One might expect the former to be most affected by access to alcohol. The latter could be considered a placebo test. You can easily reproduce Figure 4.5 in *Mastering Metrics* that combines these.

Figure is shown below. The discontinuity is evident for motor vehicle accidents but not for deaths by internal causes.



(e) Try part (c) using `rdrobust`. For comparability, use a uniform kernel and a bandwidth of 2. Compare the point estimates and standard errors to (c).

Results shown below. The point estimates of 7.66 and 9.55 correspond to those in part (c) in which we allowed the linear and quadratic terms to vary

on either side of age 21. We did not ask for the optimal bandwidth here (the manual bandwidth h=2), so rdrobust used all of the data, 24 points on each side. For comparability with (c), we did not use weighting here, but rather a uniform kernel.

```
.         rdrobust all age, c(0) h(2) kernel(uniform) p(1)

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =         48
-------------------+----------------------            BW type       =     Manual
     Number of obs |        24          24            Kernel        =    Uniform
Eff. Number of obs |        24          24            VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |     2.000       2.000
       BW bias (b) |     2.000       2.000
         rho (h/b) |     1.000       1.000

Outcome: all. Running variable: age.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  7.6627     1.3089   5.8541   0.000    5.09721      10.2282
            Robust |     -          -     4.3630   0.000    5.25872      13.8369
--------------------------------------------------------------------------------

.         rdrobust all age, c(0) h(2) kernel(uniform) p(2)

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =         48
-------------------+----------------------            BW type       =     Manual
     Number of obs |        24          24            Kernel        =    Uniform
Eff. Number of obs |        24          24            VCE method    =         NN
    Order est. (p) |         2           2
    Order bias (q) |         3           3
       BW est. (h) |     2.000       2.000
       BW bias (b) |     2.000       2.000
         rho (h/b) |     1.000       1.000

Outcome: all. Running variable: age.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  9.5478     2.1883   4.3630   0.000    5.25872      13.8369
            Robust |     -          -     3.6386   0.000    5.19606      17.3296
--------------------------------------------------------------------------------
```

(f) Use `rdbwselect` to find the optimal MSE bandwidth under different choices of polynomial (linear or quadratic) and kernel (uniform or triangular). Use these in `rdrobust` and compare your point estimates and standard errors. Normally, one could pass through the optimal bandwidth to `rdplot`, but this is problematic with these data (they are already binned).

Four optimal bandwidths are found below. In each case we asked for the bandwidth to be the same width on the left and right of age 21 (`mserd`). The four bandwidths are 0.493, 0.698, 0.451, and 0.752. These correspond to the linear and quadratic models (triangular kernel), and linear and quadratic models (uniform kernel). In the last step we put the bwselect right in the `rdrobust` command so that Stata uses the optimal bandwidth, in this case for a quadratic model with triangular kernel. The result is 8 observations per side and a discontinuity estimate of 10.28. Again, these data are already binned, which is why the resulting number of observations is so small.

```
.         rdbwselect all age, c(0) bwselect(mserd) kernel(triangular) p(1)

Bandwidth estimators for sharp RD local polynomial regression.

       Cutoff c =  | Left of c  Right of c          Number of obs =         48
-------------------+----------------------          Kernel        = Triangular
     Number of obs |        24          24          VCE method    =         NN
        Min of age |    -1.932       0.041
        Max of age |    -0.041       1.932
    Order est. (p) |         1           1
    Order bias (q) |         2           2

Outcome: all. Running variable: age.
--------------------------------------------------------------------------------
                   |         BW est. (h)         |        BW bias (b)
            Method | Left of c         Right of c | Left of c       Right of c
-------------------+-----------------------------+----------------------------
             mserd |     0.493              0.493 |     0.780            0.780
--------------------------------------------------------------------------------

.         rdbwselect all age, c(0) bwselect(mserd) kernel(triangular) p(2)

Bandwidth estimators for sharp RD local polynomial regression.

       Cutoff c =  | Left of c  Right of c          Number of obs =         48
-------------------+----------------------          Kernel        = Triangular
     Number of obs |        24          24          VCE method    =         NN
        Min of age |    -1.932       0.041
        Max of age |    -0.041       1.932
    Order est. (p) |         2           2
    Order bias (q) |         3           3

Outcome: all. Running variable: age.
--------------------------------------------------------------------------------
```

```
                   |        BW est. (h)          |       BW bias (b)
            Method | Left of c         Right of c | Left of c        Right of c
-------------------+-----------------------------+-----------------------------
             mserd |     0.698             0.698 |     0.961            0.961
--------------------------------------------------------------------------------

.         rdbwselect all age, c(0) bwselect(mserd) kernel(uniform) p(1)

Bandwidth estimators for sharp RD local polynomial regression.

       Cutoff c =  | Left of c  Right of c            Number of obs =         48
-------------------+----------------------            Kernel        =    Uniform
     Number of obs |        24          24            VCE method    =         NN
        Min of age |    -1.932       0.041
        Max of age |    -0.041       1.932
    Order est. (p) |         1           1
    Order bias (q) |         2           2

Outcome: all. Running variable: age.
--------------------------------------------------------------------------------
                   |        BW est. (h)          |       BW bias (b)
            Method | Left of c         Right of c | Left of c        Right of c
-------------------+-----------------------------+-----------------------------
             mserd |     0.451             0.451 |     0.711            0.711
--------------------------------------------------------------------------------

.         rdbwselect all age, c(0) bwselect(mserd) kernel(uniform) p(2)

Bandwidth estimators for sharp RD local polynomial regression.

       Cutoff c =  | Left of c  Right of c            Number of obs =         48
-------------------+----------------------            Kernel        =    Uniform
     Number of obs |        24          24            VCE method    =         NN
        Min of age |    -1.932       0.041
        Max of age |    -0.041       1.932
    Order est. (p) |         2           2
    Order bias (q) |         3           3

Outcome: all. Running variable: age.
--------------------------------------------------------------------------------
                   |        BW est. (h)          |       BW bias (b)
            Method | Left of c         Right of c | Left of c        Right of c
-------------------+-----------------------------+-----------------------------
             mserd |     0.752             0.752 |     1.062            1.062
--------------------------------------------------------------------------------

. //     Pass through the optimal bandwidth to rdrobust and rdplot
. //     (just using the quadratic, triangular case)
.
.         rdrobust all age, c(0) bwselect(mserd) kernel(triangular) p(2)
Estimates might be unreliable due to low number of effective observations.
```

```
Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =         48
-------------------+----------------------            BW type       =      mserd
     Number of obs |        24          24            Kernel        = Triangular
Eff. Number of obs |         8           8            VCE method    =         NN
    Order est. (p) |         2           2
    Order bias (q) |         3           3
       BW est. (h) |     0.698       0.698
       BW bias (b) |     0.961       0.961
         rho (h/b) |     0.726       0.726

Outcome: all. Running variable: age.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|     [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |   10.28     4.7765   2.1522   0.031    .918045      19.6417
            Robust |     -          -     1.9408   0.052   -.106652      21.7024
--------------------------------------------------------------------------------
```

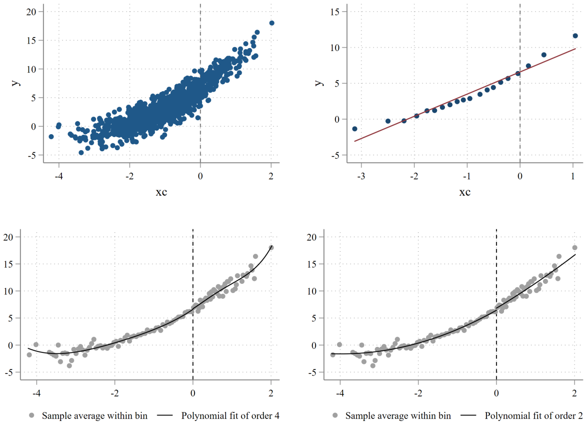
Example 2. This example will generate data with a known discontinuity in y at a threshold level of x, and then estimate a RD model. It illustrates various RD commands and results, and simulates manipulation of the running variable. (Adapted from Dale Ballou).

(a) First produce simulated data using the syntax below. Notice that x is the running variable. What is the functional relationship between the outcome y and the running variable? What is the cut score? What is the treatment effect? Is this a sharp or fuzzy regression discontinuity? Create a new variable xc that is x centered at the cut score.

```
clear
set seed 1234
drawnorm x w e u, n(1000)
gen y = 3 + 3*x + .5*x^2 + w + u
gen t = (x > 1)
replace y = y + .5*t
// centered version of x--use this below
gen xc = x-1
```

There is a quadratic relationship between the running variable x and the outcome y, as seen in line 4. The cut score is x=1 (line 5). The treatment effect is 0.5 (line 6): cases where t=1 have a value of y that is 0.5 higher than what it would be otherwise. This is a *strict* regression discontinuity since all cases where x $\leq$ 1 are untreated, and all cases where x $>$ 1 are treated.

(b) Produce a scatterplot of y against xc. Do you see evidence of a discontinuity? Try using `binscatter` and `rdplot`, the latter implementing a quartic $p = 4$ and then quadratic $p = 2$. Do you see a discontinuity in these plots? Which shows it best?



Figures shown above. It is difficult to see any discontinuity in the scatter plot. The discontinuity in the binned scatter and RD plots is evident, but slight.

(c) Estimate several (global, parameter) RD models using OLS, below. Do not use a kernel and use the full range of data. How close do these get to estimating the true treatment effect? Which model performs best, and why?

- Linear model, assuming the same slope on both sides
- Linear model, assuming different slopes on each side
- Quadratic model, assuming the same slope terms on both sides
- Quadratic model, assuming different slope terms on each side

Results below. The treatment effects in the two linear models are 2.1 and 1.1, quite a bit larger than the known effect of 0.5. The reason is that the functional relationship between y and x is misspecified. It is known to be quadratic, and we fit a linear model. The increasing slope of the relationship between y and x is mistakenly subsumed into the treatment effect. The treatment effects in the qudratic models are 0.45 and 0.51, much closer to the known effect of 0.5.

```
.         reg y xc t

      Source |       SS           df       MS      Number of obs   =     1,000
```

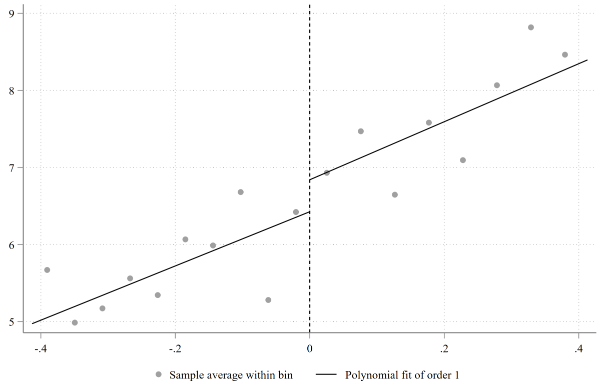
```
-------------+----------------------------------   F(2, 997)       =   2240.02
       Model |  10240.5295         2  5120.26474   Prob > F        =    0.0000
    Residual |   2278.9563       997  2.28581374   R-squared       =    0.8180
-------------+----------------------------------   Adj R-squared   =    0.8176
       Total |  12519.4858       999  12.5320178   Root MSE        =    1.5119

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          xc |   2.598992   .0622848    41.73   0.000     2.476767    2.721216
           t |   2.070868   .1711322    12.10   0.000     1.735047    2.406689
       _cons |   5.752373   .0962028    59.79   0.000      5.56359    5.941157
------------------------------------------------------------------------------

.         reg y xc t c.xc#c.xc

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(3, 996)       =   1761.62
       Model |  10534.1874         3   3511.3958   Prob > F        =    0.0000
    Residual |  1985.29838       996  1.99327146   R-squared       =    0.8414
-------------+----------------------------------   Adj R-squared   =    0.8409
       Total |  12519.4858       999  12.5320178   Root MSE        =    1.4118

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          xc |   4.033263   .1317049    30.62   0.000     3.774812    4.291714
           t |   .4473678   .2083961     2.15   0.032     .0384221    .8563135
             |
   c.xc#c.xc |   .5029089   .0414335    12.14   0.000     .4216019    .5842158
             |
       _cons |   6.429792    .105761    60.80   0.000     6.222252    6.637332
------------------------------------------------------------------------------

.         reg y c.xc##i.t

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(3, 996)       =   1612.85
       Model |  10382.3149         3  3460.77165   Prob > F        =    0.0000
    Residual |  2137.17084       996  2.14575385   R-squared       =    0.8293
-------------+----------------------------------   Adj R-squared   =    0.8288
       Total |  12519.4858       999  12.5320178   Root MSE        =    1.4648

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          xc |   2.486569   .0619109    40.16   0.000     2.365079     2.60806
         1.t |   1.132005   .2020684     5.60   0.000     .7354766    1.528534
             |
      t#c.xc |
          1  |    2.25286   .2771458     8.13   0.000     1.709003    2.796716
```

```
             |
       _cons |    5.60662   .0949179    59.07   0.000     5.420358    5.792882
------------------------------------------------------------------------------

.         reg y c.xc##c.xc##i.t

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(5, 994)       =   1055.08
       Model |  10534.5475         5  2106.90949   Prob > F        =    0.0000
    Residual |  1984.93832       994  1.99691984   R-squared       =    0.8415
-------------+----------------------------------   Adj R-squared   =    0.8407
       Total |  12519.4858       999  12.5320178   Root MSE        =    1.4131

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          xc |   4.055002     .18967    21.38   0.000     3.682803    4.427202
             |
     c.xc#c.xc |   .5096664   .0584983     8.71   0.000     .3948719    .6244608
             |
         1.t |   .3831426   .2649436     1.45   0.148    -.1367704    .9030556
             |
      t#c.xc |
          1  |   .2596287   .8106604     0.32   0.749    -1.331174    1.850431
             |
 t#c.xc#c.xc |
          1  |  -.2146489   .5198822    -0.41   0.680    -1.234842    .8055438
             |
       _cons |   6.442093     .13259    48.59   0.000     6.181905    6.702282
------------------------------------------------------------------------------
```

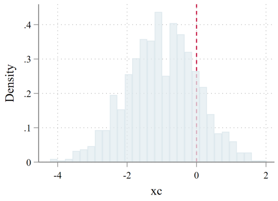
(d) Obtain local, nonparametric RD estimates using `rdrobust`. Use a linear fit, the optimal MSE bandwidth selection, and triangular kernel. Pass through the optimal bandwidth to `rdplot` to get a local RD plot. Note: the accompanying do-file also shows how to use the older command `rd`. See the help menu for that command for options.

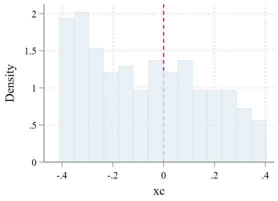
Results shown below. The point estimate using the optimal bandwidth, triangular kernel, and linear model is 0.41 with a confidence interval of (-0.52, 1.35).

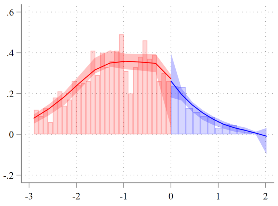
```
.         rdrobust y xc, c(0) p(1) bwselect(mserd) kernel(triangular)

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =       1000
-------------------+----------------------            BW type       =      mserd
     Number of obs |       836         164            Kernel        = Triangular
Eff. Number of obs |       129          83            VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |     0.413       0.413
       BW bias (b) |     0.658       0.658
         rho (h/b) |     0.627       0.627

Outcome: y. Running variable: xc.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  .41482     .47553   0.8723   0.383   -.517206      1.34685
            Robust |     -          -     0.9336   0.350   -.581835      1.64038
--------------------------------------------------------------------------------

.         local bandwidth = e(h_l)

.         display ‘bandwidth’
.41251756

.         // pass through bandwidth to get local RD plot
.
.         rdplot y xc if abs(xc) <= ‘bandwidth’, c(0) p(1) h(‘bandwidth’) ///
>                 kernel(triangular) graph_options(legend(position(6)) name(ballou, replace))

RD Plot with evenly spaced mimicking variance number of bins using spacings estimato rs.

         Cutoff c = 0 | Left of c  Right of c        Number of obs  =        212
----------------------+----------------------        Kernel         = Triangular
        Number of obs |       129          83
   Eff. Number of obs |       129          83
  Order poly. fit (p) |         1           1
     BW poly. fit (h) |     0.413       0.413
 Number of bins scale |     1.000       1.000

Outcome: y. Running variable: xc.
---------------------------------------------
                      | Left of c  Right of c
----------------------+----------------------
        Bins selected |        10           8
   Average bin length |     0.041       0.051
    Median bin length |     0.041       0.051
----------------------+----------------------
    IMSE-optimal bins |         7           5
  Mimicking Var. bins |        10           8
----------------------+----------------------
Rel. to IMSE-optimal: |
        Implied scale |     1.429       1.600
    WIMSE var. weight |     0.255       0.196
    WIMSE bias weight |     0.745       0.804
---------------------------------------------
```



(e) Check for manipulation in the running variable xc in two ways: by inspection using `histogram`, and using `rddensity`. Do you expect manipulation here? What does the test conclude?







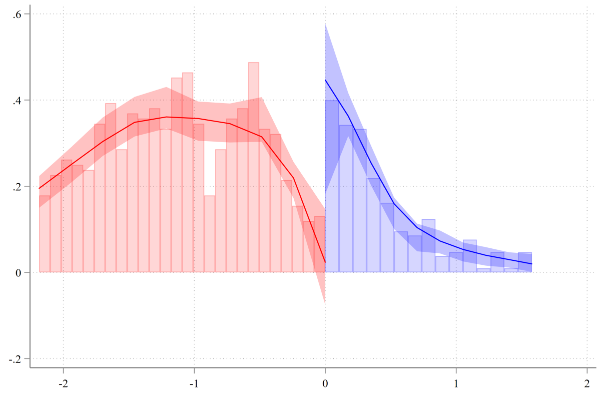
```
.         rddensity xc, c(0) plot graph_opt(name(denstest,replace) legend(off))
Computing data-driven bandwidth selectors.

RD Manipulation test using local polynomial density estimation.

     c =     0.000 | Left of c  Right of c          Number of obs =        1000
-------------------+----------------------          Model         = unrestricted
     Number of obs |       836         164          BW method     =        comb
Eff. Number of obs |       313         134          Kernel        =  triangular
    Order est. (p) |         2           2          VCE method    =   jackknife
    Order bias (q) |         3           3
       BW est. (h) |     0.966       0.858

Running variable: xc.
------------------------------------------
            Method |      T           P>|T|
-------------------+----------------------
            Robust |     1.4211      0.1553
-------------------------------------------
```

The formal density test does not detect any manipulation around the cutpoint. The test statistic is 1.4211 with a p-value of 0.1553. We cannot reject the null hypothesis of no manipulation. We would not expect any in this case, since we generated the data with no manipulation present.

(f) Now modify the data a bit to introduce manipulation in x. Try the syntax below and explain in words what the first line is doing. (Create a new centered x variable xcm) Then, re-do the `rddensity` test. Does it detect the manipulation?

```
gen xm = x
replace xm = xm + .4 if xm < 1 & xm > .65 & e > 0
gen xmc = xm-1
```

The code above is manipulating values of x between 0.65 and 1, giving them an additional 0.4 to put them over the threshold. In this case the density test is clear in showing the manipulation. The test statistic is 2.9 with a p-value of 0.0036, so we can reject the null hypothesis of no manipulation.



```
.         rddensity xmc, c(0) plot graph_opt(name(denstest2,replace) legend(off))
Computing data-driven bandwidth selectors.

RD Manipulation test using local polynomial density estimation.

     c =      0.000 | Left of c  Right of c          Number of obs =         1000
-------------------+----------------------          Model         = unrestricted
     Number of obs |       784         216          BW method     =         comb
Eff. Number of obs |       201         153          Kernel        =   triangular
    Order est. (p) |         2           2          VCE method    =    jackknife
    Order bias (q) |         3           3
       BW est. (h) |     0.728       0.526

Running variable: xmc.
------------------------------------------
            Method |      T           P>|T|
-------------------+----------------------
            Robust |    2.9102       0.0036
------------------------------------------
```

(g) Now that we know there is manipulation, try estimating the nonparametric RD using `rdrobust` (use same specs as part d). What does this yield?

For this step it is worth thinking about how y should be changed, if at all. If we assume that cases manipulated into the treatment group get the same effect from being exposed to the treatment, then we should add the 0.5 to these cases (as below). One could also leave the original y's intact, but this would be assuming no treatment effect for these manipulated cases. The RD estimate of the treatment effect is too small in this case: -0.23. This is likely because cases that would have (without manipulation) not been treated were in fact treated. These cases had somewhat lower potential outcomes than other non-manipulated cases just above the cutoff.

```
.         gen y2=y

.         replace y2=y+0.5 if (x<1 & x>0.65 & e>0)
(52 real changes made)

.         rdrobust y2 xmc, c(0) p(1) bwselect(mserd) kernel(triangular)

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =       1000
-------------------+----------------------            BW type       =      mserd
     Number of obs |       784         216            Kernel        = Triangular
Eff. Number of obs |        44         111            VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |     0.308       0.308
       BW bias (b) |     0.512       0.512
         rho (h/b) |     0.602       0.602

Outcome: y2. Running variable: xmc.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z      P>|z|     [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional | -.22507     .53709    -0.4191  0.675    -1.27774         .8276
            Robust |     -          -      -0.0206  0.984    -1.2555      1.22936
--------------------------------------------------------------------------------
```

Example 3. This example, based on an example created by Celeste Carruthers, also uses simulated data to estimate the effect of participation in a gifted and talented (G&T) program.

(a) Generate 10,000 student observations. The data will include a measure of students' "true ability," *trueability* $\sim N(50, 4)$, and their 3rd grade test score, which is a noisy measure of their true ability $grade3test = trueability + u$ where $u \sim N(0, 1)$. To add a bit of realism, we will round test scores to the nearest 0.25 to create a discrete scale.

```
clear
set seed 195423
set obs 10000
gen id=_n
gen trueability = 50 + 4*rnormal()
gen grade3test  = trueability + rnormal()
replace grade3test = round(grade3test, 0.25)
```

(b) Suppose 3rd graders scoring at or above 56 are eligible for the G&T program. Create a treatment assignment variable re-centered at zero, and a "gap" variable that contains the distance between the running variable and the cut score.

```
gen above56 = (grade3test>=56)
gen gap = grade3test-56
```

(c) Assume perfect compliance. Create an indicator variable for G&T participation *inGT*, that equals one for treated students and zero otherwise. What proportion of students are treated? Try estimating an OLS regression for G&T participation where *inGT* is regressed on the *gap* and the threshold indicator *above56*. What happens and why?

7.59% of students participated in G&T. When you regress the treatment *inGT* on the *gap* and threshold indicator *above56*, Stata cannot produce estimates. This is because *above56* perfectly determines the outcome *inGT*. (It is a strict, not fuzzy, discontinuity).

```
. gen inGT=(above56==1)

. sum inGT

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        inGT |     10,000       .0759    .2648513          0          1

. reg inGT gap above56

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      =         .
       Model |   701.3919         2  350.69595   Prob > F        =         .
    Residual |          0     9,997           0   R-squared       =    1.0000
-------------+----------------------------------   Adj R-squared   =    1.0000
       Total |   701.3919     9,999  .070146205   Root MSE        =         0

------------------------------------------------------------------------------
        inGT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gap |          0  (omitted)
     above56 |          1          .        .       .            .           .
       _cons |          0  (omitted)
------------------------------------------------------------------------------
```

(d) Create an outcome variable (*grade 4* test score) such that G&T participation has a positive treatment effect of +3 points. Assume that test growth from 3rd to 4th grade would be 5 points in the absence of treatment. As before, we will include some random noise, and round the test scale to the nearest 0.25.

```
gen grade4test = round(trueability + 5 + rnormal() + (3*inGT), 0.25)
```

(e) Estimate an RD model for 4th grade test scores, assuming a linear relationship with the running variable (3rd grade test scores). Use the full range of data and no kernel. First do this assuming the same slope on either side of the cut score. Then, allow the slope to vary on either side. Is there evidence of a change in slope beyond the cut score? Does this finding make sense to you? Repeat using `rdrobust` with the default options.

Results below. There is no evidence of a change in slope beyond the cut score. This makes sense since we know the original data generating process, in which *grade4test* is linear in the running variable. The estimated discontinuity is a +3.02 increase in grade 4 test scores. The rdrobust point estimate is 3.22.

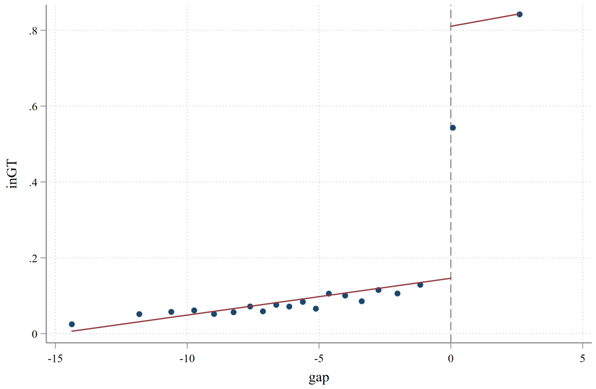
```
.         reg grade4test gap i.above56

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      =  48332.69
       Model |  187161.691         2  93580.8457   Prob > F        =    0.0000
    Residual |  19356.0026     9,997  1.93618111   R-squared       =    0.9063
-------------+----------------------------------   Adj R-squared   =    0.9063
       Total |  206517.694     9,999  20.6538348   Root MSE        =    1.3915

------------------------------------------------------------------------------
  grade4test |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gap |   .9392992   .0040454   232.19   0.000     .9313694     .947229
   1.above56 |   3.029696   .0624409    48.52   0.000       2.9073    3.152093
       _cons |   60.61507   .0304137  1993.02   0.000     60.55545    60.67468
------------------------------------------------------------------------------
```

```
.         reg grade4test c.gap##i.above56

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(3, 9996)      =  32218.65
       Model |  187161.733         3  62387.2445   Prob > F        =    0.0000
    Residual |  19355.9606     9,996   1.9363706   R-squared       =    0.9063
-------------+----------------------------------   Adj R-squared   =    0.9062
       Total |  206517.694     9,999  20.6538348   Root MSE        =    1.3915

------------------------------------------------------------------------------
  grade4test |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gap |    .939221   .0040802   230.19   0.000     .9312229    .9472191
   1.above56 |    3.02236   .0798731    37.84   0.000     2.865792    3.178927
             |
above56#c.gap |
          1  |   .0046223   .0313781     0.15   0.883    -.0568851    .0661296
             |
       _cons |   60.61455   .0306168  1979.78   0.000     60.55453    60.67456
------------------------------------------------------------------------------

.         rdrobust grade4test gap, c(0) p(1) kernel(triangular)
Mass points detected in the running variable.

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =      10000
-------------------+----------------------            BW type       =      mserd
     Number of obs |      9241         759            Kernel        = Triangular
Eff. Number of obs |       928         525            VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |     2.190       2.190
       BW bias (b) |     3.633       3.633
         rho (h/b) |     0.603       0.603
        Unique obs |        80          33

Outcome: grade4test. Running variable: gap.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  3.2231     .15821   20.3725  0.000    2.91305      3.53322
            Robust |     -          -     17.4736  0.000    2.90155      3.63471
--------------------------------------------------------------------------------
Estimates adjusted for mass points in the running variable.
```

(f) Drop the existing *inGT* and *grade4test* variables and re-create them assuming a “fuzzy” GT treatment that increases smoothly with grade 3 test scores and then jumps discontinuously (by about 70 percentage points) at the cut score. This might arise if G&T placement is dependent on the grade 3 test score as well as other factors (e.g., parental input, teacher recommendation). Use the syntax below. Now what proportion of students are treated, overall? Below the cutoff? Above? Use `binscatter` to visualize the relationship between treatment and the grade 3 score.

```
drop inGT grade4test
gen inGT=round(-.77+.007*grade3test+0.7*above56+runiform())
gen grade4test = round(trueability + 5 + rnormal() + (3*inGT), 0.25)
```

13.3% of students are treated, overall. Below the cutoff, 7.6% of students are selected for the G&T program. Above the cutoff, 83.4% are selected. The discontinuous jump in treatment status is seen in the binscatter below.



(g) As in (c), estimate a regression for G&T placement where *inGT* is regressed on the *gap* and the threshold indicator *above56*. Interpret your results. (Try estimating this in two ways: first assuming the slope is constant on either side of the cutoff, and then allowing the slope to change). For later use, use the `predict` command to get predicted values for treatment (placement in G&T) given the 3rd grade score. Call this variable *predGT*.

Results below. The first regression tells us that the probability of selection for G&T increases with *gap* (the student's score minus 56). There is also a discontinuous jump in the probability of selection at the cut score, of 70.3 percentage points.

```
.         reg inGT gap i.above56

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      =   2730.15
       Model |  408.373259         2  204.186629   Prob > F        =    0.0000
    Residual |  747.671141     9,997  .074789551   R-squared       =    0.3533
-------------+----------------------------------   Adj R-squared   =    0.3531
       Total |   1156.0444     9,999  .115616002   Root MSE        =    .27348
```

```
------------------------------------------------------------------------------
        inGT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gap |   .0066524   .0007951     8.37   0.000     .0050939    .0082109
   1.above56 |   .7026519    .012272    57.26   0.000     .6785962    .7267075
       _cons |   .1198432   .0059775    20.05   0.000     .1081262    .1315602
------------------------------------------------------------------------------

.         reg inGT c.gap##i.above56

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(3, 9996)      =   1820.23
       Model |  408.418692         3  136.139564   Prob > F        =    0.0000
    Residual |  747.625708     9,996  .074792488   R-squared       =    0.3533
-------------+----------------------------------   Adj R-squared   =    0.3531
       Total |   1156.0444     9,999  .115616002   Root MSE        =    .27348

-------------------------------------------------------------------------------
         inGT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
          gap |   .0065711   .0008019     8.19   0.000     .0049992     .008143
    1.above56 |    .695023   .0156977    44.28   0.000     .6642524    .7257936
              |
above56#c.gap |
           1  |   .0048064   .0061668     0.78   0.436    -.0072818    .0168946
              |
        _cons |   .1193058   .0060172    19.83   0.000     .1075109    .1311008
-------------------------------------------------------------------------------

.         // get predicted treatment for later use
.         predict predGT
(option xb assumed; fitted values)
```

(h) Re-estimate the RD models from part (e) assuming a linear relationship with the running variable. Assume the discontinuity is “sharp,” even though we know otherwise. How does the estimated treatment effect differ from the known treatment effect of 3 points?

Results below. The estimated treatment effect is smaller, at 2.0 versus the known 3 points. This is not surprising: when the discontinuity is fuzzy, the difference in outcomes around the cutoff will be smaller, since not all above the cut score were treated, and some of those below the cut score were treated.

```
.         reg grade4test gap i.above56

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      =  35035.27
       Model |  180812.138         2  90406.0691   Prob > F        =    0.0000
    Residual |  25796.5606     9,997  2.58043019   R-squared       =    0.8751
```

```
-------------+----------------------------------   Adj R-squared   =    0.8751
       Total |  206608.699     9,999  20.6629362   Root MSE        =    1.6064

------------------------------------------------------------------------------
  grade4test |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gap |   .9617267   .0046702   205.93   0.000     .9525722    .9708812
   1.above56 |    2.06733   .0720845    28.68   0.000      1.92603    2.208631
       _cons |    61.0054   .0351109  1737.50   0.000     60.93658    61.07423
------------------------------------------------------------------------------

.         reg grade4test c.gap##i.above56

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(3, 9996)      =  23357.13
       Model |  180814.673         3  60271.5578   Prob > F        =    0.0000
    Residual |  25794.0253     9,996   2.5804347   R-squared       =    0.8752
-------------+----------------------------------   Adj R-squared   =    0.8751
       Total |  206608.699     9,999  20.6629362   Root MSE        =    1.6064

-------------------------------------------------------------------------------
   grade4test |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
          gap |   .9611196   .0047102   204.05   0.000     .9518867    .9703525
    1.above56 |   2.010342   .0922047    21.80   0.000     1.829602    2.191082
              |
above56#c.gap |
           1  |   .0359043   .0362225     0.99   0.322    -.0350991    .1069077
              |
        _cons |   61.00139   .0353438  1725.95   0.000     60.93211    61.07067
-------------------------------------------------------------------------------

.         rdrobust grade4test gap, c(0) p(1) kernel(triangular)
Mass points detected in the running variable.

Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =      10000
-------------------+----------------------            BW type       =      mserd
     Number of obs |      9241         759            Kernel        = Triangular
Eff. Number of obs |       781         487            VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |     1.995       1.995
       BW bias (b) |     3.275       3.275
         rho (h/b) |     0.609       0.609
        Unique obs |        80          33

Outcome: grade4test. Running variable: gap.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
```

```
-------------------+------------------------------------------------------------
      Conventional |  2.2697     .20845    10.8880  0.000     1.86109      2.67822
            Robust |     -          -      9.4841   0.000     1.84767      2.81027
--------------------------------------------------------------------------------
Estimates adjusted for mass points in the running variable.
```

(i) We will now estimate the treatment effect using RD but allowing for non-compliance (fuzzy RD). Do this two ways:

- Use `rdrobust` with the default options but add the option `fuzzy(inGT)`.
- Using two stage least squares. One way to implement this is to regress *grade4test* on the predicted *inGT* from (g). `ivregress` is a little trickier in this context—you'll need to manually create "gapabove" and "gapbelow" variables if you want the slope to vary above and below *c*. See the syntax below.

```
gen gapabove = gap*above56
gen gapbelow = gap*(1-above56)
ivregress 2sls grade4test (inGT=above56) gapbelow gapabove , first
```

Note the `rdrobust` and manual estimates are not directly comparable since the former uses the optimal bandwidth. How close do these get to the "true" treatment effect of 3 points?

See below. rdrobust with fuzzy recovers the known treatment effect of +3. The 2SLS and ivregress approaches also yield point estimates close to 3.

```
.         // first using rdrobust with default options
.         rdrobust grade4test gap, c(0) p(1) kernel(triangular) fuzzy(inGT)
Mass points detected in the running variable.

Fuzzy RD estimates using local polynomial regression.

      Cutoff c = 0 | Left of c  Right of c            Number of obs =      10000
-------------------+----------------------            BW type       =      mserd
     Number of obs |      9241         759            Kernel        = Triangular
Eff. Number of obs |       928         525            VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |     2.188       2.188
       BW bias (b) |     4.152       4.152
         rho (h/b) |     0.527       0.527
        Unique obs |        80          33

First-stage estimates. Outcome: inGT. Running variable: gap.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
```

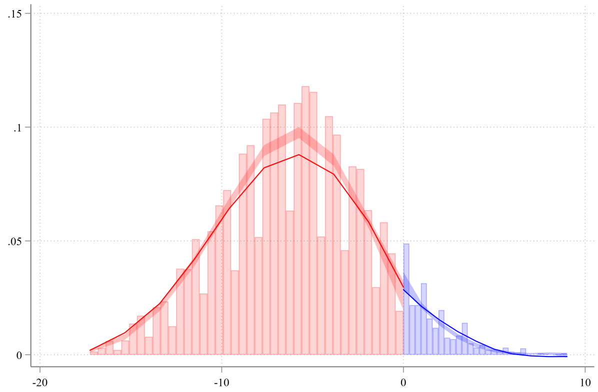
```
      Conventional |  .74416     .03788   19.6451  0.000     .66992      .818409
            Robust |     -          -     17.9063  0.000    .674327       .84009
--------------------------------------------------------------------------------

Treatment effect estimates. Outcome: grade4test. Running variable: gap. Treatment St
> atus: inGT.
--------------------------------------------------------------------------------
            Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
-------------------+------------------------------------------------------------
      Conventional |  3.0414     .21207   14.3415  0.000    2.62573      3.45702
            Robust |     -          -     12.7669  0.000    2.58462      3.52212
--------------------------------------------------------------------------------
Estimates adjusted for mass points in the running variable.

.         // then 2SLS (second stage shown here using predicted GT)
.         reg grade4test gap predGT

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      =  35037.76
       Model |  180813.742         2  90406.8711   Prob > F        =    0.0000
    Residual |  25794.9566     9,997  2.58026974   R-squared       =    0.8752
-------------+----------------------------------   Adj R-squared   =    0.8751
       Total |  206608.699     9,999  20.6629362   Root MSE        =    1.6063

------------------------------------------------------------------------------
  grade4test |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gap |    .942128   .0050712   185.78   0.000     .9321874    .9520686
      predGT |   2.943022   .1025765    28.69   0.000     2.741951    3.144093
       _cons |   60.65253   .0434644  1395.45   0.000     60.56734    60.73773
------------------------------------------------------------------------------

.         // ivregress is tricky here since "above56" is an exogenous instrument
.         // conditional on grade3test. Can't use above56 both as the instrument
.         // and in the interaction with "gap". So manually created an interaction
.         // below
.
.         gen gapabove = gap*above56

.         gen gapbelow = gap*(1-above56)

.         // note gapbelow and gapabove estimate slope on gap below and above
.         ivregress 2sls grade4test (inGT=above56) gapbelow gapabove , first

First-stage regressions
-----------------------

                                                Number of obs     =     10,000
                                                F(   3,   9996)   =    1820.23
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3533
                                                Adj R-squared     =     0.3531
                                                Root MSE          =     0.2735

------------------------------------------------------------------------------
        inGT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gapbelow |   .0065711   .0008019     8.19   0.000     .0049992     .008143
    gapabove |   .0113775   .0061145     1.86   0.063    -.0006081    .0233631
     above56 |    .695023   .0156977    44.28   0.000     .6642524    .7257936
       _cons |   .1193058   .0060172    19.83   0.000     .1075109    .1311008
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression          Number of obs   =     10,000
                                                  Wald chi2(3)    =   93654.38
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.9066
                                                  Root MSE        =     1.3895

------------------------------------------------------------------------------
  grade4test |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        inGT |   2.892483   .1147519    25.21   0.000     2.667573    3.117392
    gapbelow |   .9421128   .0043867   214.77   0.000      .933515    .9507106
    gapabove |   .9641147   .0319592    30.17   0.000     .9014757    1.026754
       _cons |    60.6563   .0379859  1596.81   0.000     60.58185    60.73075
------------------------------------------------------------------------------
Instrumented:  inGT
Instruments:   gapbelow gapabove above56
```

(j) Finally, do a manipulation test using `rddensity`. What does it find? Would you expect to see evidence of manipulation here?

Results are below. We cannot reject the null hypothesis of no manipulation. This is the correct decision, since we know from generating the data that 3rd grade test scores were not manipulated.

```
.         rddensity gap, c(0) plot graph_opt(name(fuzzy, replace) legend(off))
Computing data-driven bandwidth selectors.

Point estimates and standard errors have been adjusted for repeated observations.
(Use option nomasspoints to suppress this adjustment.)

RD Manipulation test using local polynomial density estimation.

     c =     0.000 | Left of c  Right of c          Number of obs =        10000
-------------------+----------------------          Model         = unrestricted
     Number of obs |      9241         759          BW method     =         comb
Eff. Number of obs |      4153         737          Kernel        =   triangular
    Order est. (p) |         2           2          VCE method    =    jackknife
    Order bias (q) |         3           3
```

```
       BW est. (h) |     5.750       5.750

Running variable: gap.
------------------------------------------
            Method |      T           P>|T|
-------------------+----------------------
            Robust |    1.5199      0.1285
------------------------------------------
```


.15
.1
.05
0
-20
-10
0
10