`regsensitivity`: A Stata Package for Regression Sensitivity Analysis

Paul Diegert* Matthew A. Masten† Alexandre Poirier‡

June 7, 2022

Introduction

Omitted variables are one of the most important threats to the identification of causal effects. In linear models, the well known omitted variable bias formula shows how an omitted variable can bias the regression coefficient on the covariate of interest when that covariate is correlated with the omitted variable. Since is often implausible to assume that data has been collected on every relevant variable, applied research is often vulernable to this bias. Nonetheless, omitted variable bias can be quantified under various alternative assumptions about the relationship between the omitted variable and the covariate of interest. Using these techniques, researchers can analyze how sensitive their results are to omitted variable bias.

Several methods of sensitivity analysis for linear models have been proposed in the literature. The `regsensitivity` package implements the methods proposed in Diegert, Masten, and Poirier (2022). In the paper, the authors define a set of sensitivity parameters which index relaxations of the assumption that the covariate of interest is uncorrelated with any unobserved variables. The parameter of interest in both cases is β_{long}, the coefficient on that covariate of interest in the infeasible regression that includes the unobserved variables. Using this framework, we can ask two questions:

1. What is the set of parameter estimates for β_{long} which are consistent with the relaxed assumptions? That is, what are bounds on the value of β_{long} under the alternate assumptions?
2. How much can we relax the exogeneity assumption before a hypothesis about β_{long} is overturned? This is called the *breakdown point*: the maximum relaxation of the baseline assumption before the hypothesis is overturned.

`regsensitivity` can be used to perform both of these sensitivity analyses.

Getting Started

We will illustrate how to use `regsensitivity` with data from Bazzi, Fiszbein, and Gebresilasse (2020), which is used in the empirical application in Diegert, Masten, and Poirier (2022). One of the datasets used in Bazzi, Fiszbein, and Gebresilasse (2020) is included with the package, and can be loaded using the `sysuse` command:

```
. sysuse bfg2020, clear
```

*Department of Economics, Duke University, paul.diegert@duke.edu
†Department of Economics, Duke University, matt.masten@duke.edu
‡Department of Economics, Georgetown Univervsity, alexandre.poirier@georgetown.edu

The specification in column (7) of Table III in Bazzi, Fiszbein, and Gebresilasse (2020) and replicated in Diegert, Masten, and Poirier (2022) is as follows,

```
. local y avgrep2000to2016

. local x tye_tfe890_500kNI_100_l6

. local w1 log_area_2010 lat lon temp_mean rain_mean elev_mean d_coa d_riv d_lak ave_gyi

. local w0 i.statea

. local w `w1´ `w0´

. local SE cluster(km_grid_cel_code)

. reg `y´ `x´ `w´, `SE´

Linear regression                               Number of obs     =      2,036
                                                F(39, 379)        =          .
                                                Prob > F          =          .
                                                R-squared         =     0.3321
                                                Root MSE          =     9.6368

                                 (Std. err. adjusted for 380 clusters in km_grid_cel_code)
```

avgrep2000to2016	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
tye_tfe890_500kNI_100_l6	2.054759	.3491648	5.88	0.000	1.368217	2.741302
log_area_2010	.2758775	.979906	0.28	0.778	-1.650856	2.202611
lat	2.26515	1.101151	2.06	0.040	.1000189	4.430281
lon	.0108189	.2913783	0.04	0.970	-.5621017	.5837395
temp_mean	1.62737	1.068132	1.52	0.128	-.4728361	3.727577
rain_mean	.0164826	.0046086	3.58	0.000	.007421	.0255442
elev_mean	.0154764	.0037786	4.10	0.000	.0080468	.022906
d_coa	9.83e-06	3.76e-06	2.62	0.009	2.45e-06	.0000172
d_riv	.0000307	9.91e-06	3.10	0.002	.0000112	.0000502
d_lak	3.05e-07	4.45e-06	0.07	0.945	-8.44e-06	9.05e-06
ave_gyi	-3.779807	10.81002	-0.35	0.727	-25.03493	17.47532
statea						
5	-4.213545	3.386398	-1.24	0.214	-10.87203	2.444936
8	-27.31682	6.246914	-4.37	0.000	-39.59977	-15.03387
12	4.627587	3.354655	1.38	0.169	-1.968479	11.22365
13	.5398875	2.643504	0.20	0.838	-4.657883	5.737658
17	-10.25822	3.787414	-2.71	0.007	-17.7052	-2.811248
18	-5.924452	3.497393	-1.69	0.091	-12.80118	.9522727
19	-18.02705	4.514016	-3.99	0.000	-26.9027	-9.151398
20	1.598741	4.633634	0.35	0.730	-7.512109	10.70959
21	-.504168	3.185907	-0.16	0.874	-6.768436	5.7601
22	.8939823	3.276872	0.27	0.785	-5.549145	7.337109
26	-14.06314	4.40552	-3.19	0.002	-22.72546	-5.400816
27	-18.10308	4.821495	-3.75	0.000	-27.58331	-8.622851
28	-6.930918	3.675983	-1.89	0.060	-14.15879	.2969573
29	-4.170334	3.902039	-1.07	0.286	-11.84269	3.502024
31	-1.342615	4.73751	-0.28	0.777	-10.65771	7.97248
35	-40.78007	9.264248	-4.40	0.000	-58.99583	-22.56431
36	-9.821649	4.68884	-2.09	0.037	-19.04105	-.6022507
37	-13.53756	4.241671	-3.19	0.002	-21.87772	-5.197404
38	-11.98193	5.512474	-2.17	0.030	-22.82079	-1.143061
39	-6.190808	3.655508	-1.69	0.091	-13.37843	.9968088
40	13.60029	4.880539	2.79	0.006	4.003963	23.19661
42	-3.14623	4.426406	-0.71	0.478	-11.84962	5.557161
46	-11.84706	5.101547	-2.32	0.021	-21.87794	-1.816175
47	-3.541445	2.794141	-1.27	0.206	-9.035406	1.952515
48	12.82591	4.174157	3.07	0.002	4.618502	21.03331
51	-.8116892	4.047756	-0.20	0.841	-8.770561	7.147182
54	-3.243583	3.570236	-0.91	0.364	-10.26353	3.776369
55	-18.92918	4.503985	-4.20	0.000	-27.78511	-10.07325
56	-19.02288	9.729311	-1.96	0.051	-38.15307	.1073112

```
               _cons |  -73.53523   57.84708    -1.27   0.204    -187.2766    40.20618
```

To run the default sensitivity analysis, simply run,

```
. regsensitivity `y´ `x´ `w´, compare(`w1´)

Regression Sensitivity Analysis, Bounds

Analysis         : DMP (2022)                 Number of obs      =      2,036
                                              Beta(short)        =      1.925
Treatment        : tye_tfe890_500kNI_100_l6   Beta(medium)       =      2.055
Outcome          : avgrep2000to2016           R2(short)          =      0.033
                                              R2(medium)         =      0.105
                                              Var(Y)             =    101.739
                                              Var(X)             =      0.901
                                              Var(X_Residual)    =      0.882

Hypothesis       : Beta > 0                   Breakdown point    =      80.4%
Other Params     : cbar = 1, rybar = +inf
--------------------------------------------------------------------------------
 rxbar                             Beta
--------------------------------------------------------------------------------
 0.000                            [  2.05,    2.05 ]
 0.095                            [  1.91,    2.20 ]
 0.196                            [  1.76,    2.35 ]
 0.296                            [  1.59,    2.52 ]
 0.397                            [  1.41,    2.70 ]
 0.497                            [  1.20,    2.91 ]
 0.592                            [  0.95,    3.16 ]
 0.693                            [  0.61,    3.50 ]
 0.793                            [  0.07,    4.04 ]
 0.894                            [ -1.05,    5.16 ]
 0.989                            [-58.90,   63.01 ]
 0.989                            [  -inf,   +inf  ]
--------------------------------------------------------------------------------

.
```

The output shows results answers to each of the questions mentioned in the introduction: what are the bounds on β_{long} under a range of assumptions, and at what point does the hypothesis $\beta_{\text{long}} > 0$ break down. To explore the output and capibilities of the package in more detail we consider each of the analyses separately.

Bounds

Diegert, Masten, and Poirier (2022) consider the model:

$$Y = \beta_{\text{long}} X + \gamma_0' W_0 + \gamma_1' W_1 + \gamma_2 W_2 + Y^{\perp X,W},$$

where (Y, X, W_0, W_1) are observed and W_2 is an omitted variable that is potentially correlated with (X, W_0, W_1).[1] Restrictions on the joint distribution of (X, W_0, W_1, W_2) are governed by three scalar sensitivity parameters, $(\bar{r}_X, \bar{r}_Y, \bar{c})$. Given the joint distribution of the observed variables, (Y, X, W_0, W_1), and the values of the sensitivity parameters, Diegert, Masten, and Poirier (2022) show how to compute the upper and lower bounds on the identified set for β_{long}, denoted by $\mathcal{B}_I(\bar{r}_X, \bar{r}_Y, \bar{c})$. The identified set is the set of values of β_{long} which are consistent with the distribution of observed data and the maintained assumptions. When $\bar{r}_X > 0$ and $\bar{r}_Y > 0$, β_{long} is not point identified, so we instead estimate these bounds. For more details about the definitions and interpretation of the sensitivity parameters, see Diegert, Masten, and Poirier (2022).

[1] We denote the coefficient on X by β_{long} because it is the regression coefficient in the infeasible "long" regression of Y on $(1, X, W_0, W_1, W_2)$. This helps distinguish it from β_{med}, the coefficient on X in the regression of Y on $(1, X, W_0, W_1)$, and from β_{short}, the coefficient on X in the regression of Y on $(1, X, W_0)$.

`regsensitivity` can be used with the option `bounds` subcommand to calculate the upper and lower bounds of $\mathcal{B}_I(\bar{r}_X, \bar{r}_Y, \bar{c})$. The basic syntax for `regsensitivity` is similar to the `regress` command and its variants:

`regsensitivity bounds` *depvar indepvar controls*, *options...*,

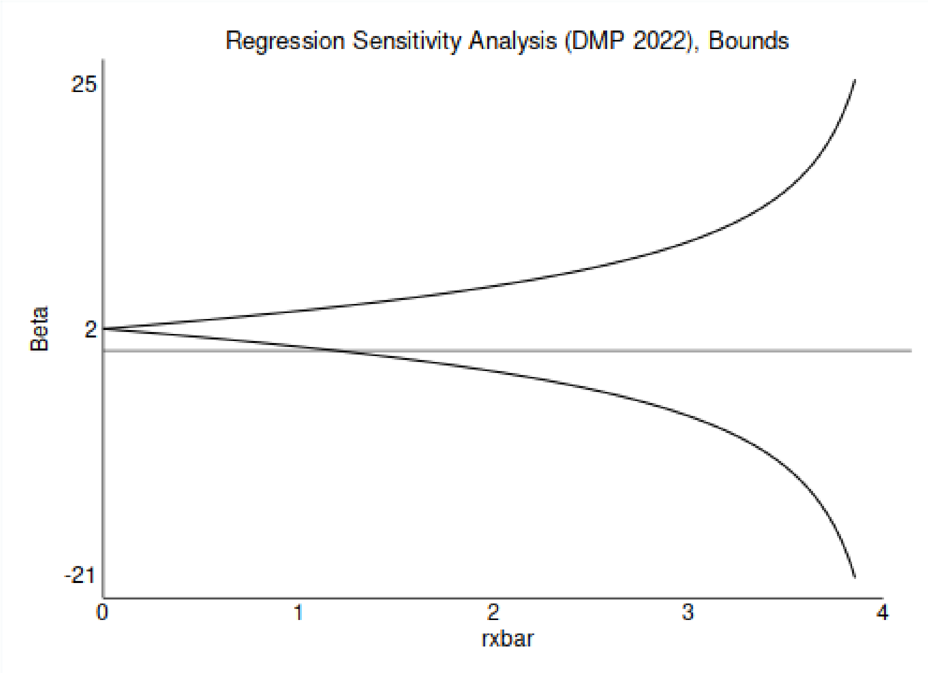
where *depvar* is the dependant variable, Y, and *indepvar controls* are the independent variables, (X, W_0, W_1). Unlike `regress`, the order of the independent variables matter in the call to `regsensitivity`. The first variable, *indepvar*, is X, the variable of interest for which the sensitivity analysis is conducted while *controls* are additional variables included in the model which are not of interest.

By default, `regsensitivity bounds` calculates the bounds for a range of values of $\bar{r}_X$ holding $\bar{c}$ and $\bar{r}_Y$ fixed. The defaults are to set $\bar{c} = 1$ and $\bar{r}_Y = +\infty$. To specify a different value of $\bar{c}$, use the `cbar` option. For example,

```
. regsensitivity bounds `y´ `x´ `w´, compare(`w1´) cbar(.1)

Regression Sensitivity Analysis, Bounds

Analysis         : DMP (2022)                Number of obs      =      2,036
                                             Beta(short)        =      1.925
Treatment        : tye_tfe890_500kNI_100_l6  Beta(medium)       =      2.055
Outcome          : avgrep2000to2016          R2(short)          =      0.033
                                             R2(medium)         =      0.105
                                             Var(Y)             =    101.739
                                             Var(X)             =      0.901
                                             Var(X_Residual)    =      0.882

Hypothesis       : Beta > 0                  Breakdown point    =       119%
Other Params     : cbar = .1, rybar = +inf
--------------------------------------------------------------------------------
 rxbar                              Beta
--------------------------------------------------------------------------------
 0.000                             [  2.05,   2.05 ]
 0.394                             [  1.45,   2.66 ]
 0.808                             [  0.74,   3.37 ]
 1.202                             [ -0.02,   4.13 ]
 1.617                             [ -0.93,   5.04 ]
 2.032                             [ -2.02,   6.13 ]
 2.425                             [ -3.32,   7.43 ]
 2.840                             [ -5.17,   9.28 ]
 3.234                             [ -7.86,  11.97 ]
 3.648                             [-13.63,  17.74 ]
 4.042                             [-75.24,  79.35 ]
 4.063                             [  -inf,   +inf ]
--------------------------------------------------------------------------------
```

To plot the results, use the `plot` subcommand,

```
. regsensitivity plot
```



Notice that in the call to **regsensitivity bounds**, we also included an option compare(*varlist*). This specifies which of the variables in the *controls* are included in W_1 rather than W_0. These are referred to as the *comparison controls* because they are the variables used to calibrate the sensitivity parameters, $(\bar{r}_X, \bar{r}_Y, \bar{c})$. For more details, see section 3.3 in Diegert, Masten, and Poirier (2022).

By including more variables in the comparison controls, the identified set will tend to be larger for a given value of the sensitivity parameters. For example, if the compare option is omitted, then all the control variables are included in W_1,

```
. regsensitivity bounds `y´ `x´ `w´, cbar(.1)

Regression Sensitivity Analysis, Bounds

Analysis         : DMP (2022)                Number of obs      =      2,036
                                             Beta(short)        =      1.708
Treatment        : tye_tfe890_500kNI_100_16  Beta(medium)       =      2.055
Outcome          : avgrep2000to2016          R2(short)          =      0.027
                                             R2(medium)         =      0.332
                                             Var(Y)             =    136.320
                                             Var(X)             =      1.257
                                             Var(X_Residual)    =      0.882

Hypothesis       : Beta > 0                  Breakdown point    =       29.7%
Other Params     : cbar = .1, rybar = +inf
--------------------------------------------------------------------------------
 rxbar                             Beta
--------------------------------------------------------------------------------
 0.000                            [   2.05,     2.05 ]
 0.128                            [   1.20,     2.91 ]
 0.262                            [   0.25,     3.86 ]
 0.396                            [  -0.77,     4.88 ]
 0.531                            [  -1.91,     6.02 ]
 0.665                            [  -3.24,     7.35 ]
 0.800                            [  -4.88,     8.99 ]
 0.934                            [  -7.07,    11.18 ]
 1.069                            [ -10.45,    14.56 ]
 1.203                            [ -17.46,    21.57 ]
```

```
1.331                               [-106.48,  110.59 ]
1.336                               [   -inf,    +inf ]
```

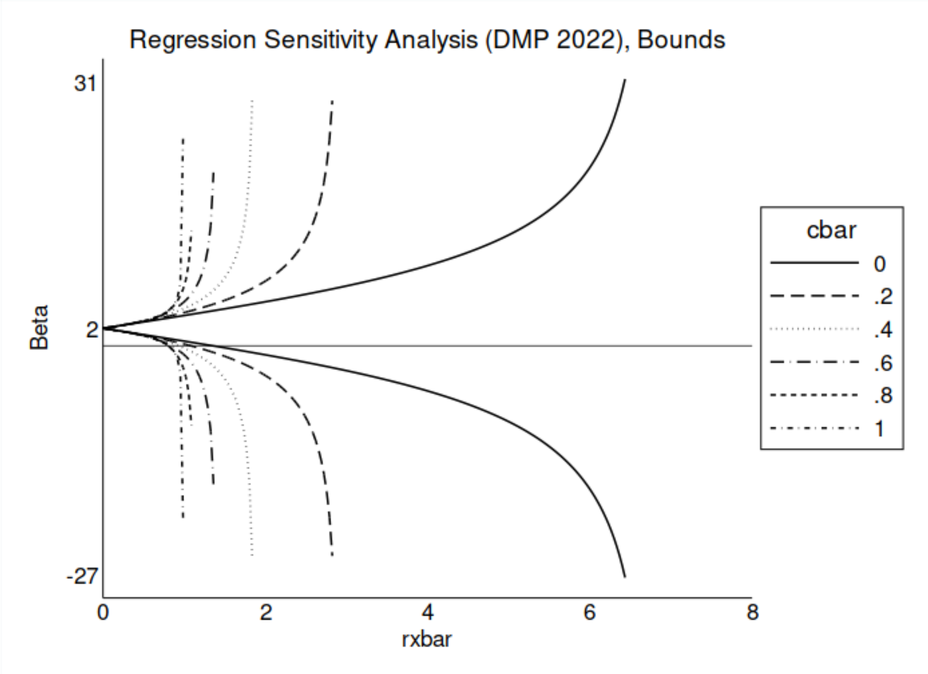
With all the controls included in W_1, the identified set becomes $\mathbb{R}$ at $\bar{r}_X = 1.336$, compared to $\bar{r}_X = 4.063$ when W_1 excludes the state fixed effects (*statea*). To directly compare the bounds under the two choices of W_1, we can manually set the values of `rxbar` to be the same in each case. The output table from the last call to `regsensitivity bounds` are stored in `e(idset_table)`. We can extract the values of `rxbar` from these and rerun the analysis with W_1 excluding state fixed effects as follows,

```
. forvalues i=1/12{
  2.     local rxbar `rxbar´ `=e(idset_table)[`i´, 1]´
  3. }

. regsensitivity bounds `y´ `x´ `w´, compare(`w1´) cbar(.1) rxbar(`rxbar´)

Regression Sensitivity Analysis, Bounds

Analysis        : DMP (2022)                Number of obs      =      2,036
                                            Beta(short)        =      1.925
Treatment       : tye_tfe890_500kNI_100_l6  Beta(medium)       =      2.055
Outcome         : avgrep2000to2016          R2(short)          =      0.033
                                            R2(medium)         =      0.105
                                            Var(Y)             =    101.739
                                            Var(X)             =      0.901
                                            Var(X_Residual)    =      0.882

Hypothesis      : Beta > 0                  Breakdown point    =        119%
Other Params    : cbar = .1, rybar = +inf
--------------------------------------------------------------------------------
 rxbar                               Beta
--------------------------------------------------------------------------------
 0.000                              [ 2.0548,  2.0548 ]
 0.128                              [ 1.8628,  2.2468 ]
 0.262                              [ 1.6550,  2.4546 ]
 0.396                              [ 1.4408,  2.6687 ]
 0.531                              [ 1.2198,  2.8898 ]
 0.665                              [ 0.9911,  3.1184 ]
 0.800                              [ 0.7540,  3.3555 ]
 0.934                              [ 0.5078,  3.6017 ]
 1.069                              [ 0.2513,  3.8583 ]
 1.203                              [-0.0166,  4.1262 ]
 1.331                              [-0.2829,  4.3924 ]
 1.336                              [-0.2937,  4.4032 ]
--------------------------------------------------------------------------------
```

Comparing each line of the table to the previous call where all the *controls* were included in W_1, we can see that the bounds are much tigher for each value of $\bar{r}_X$.

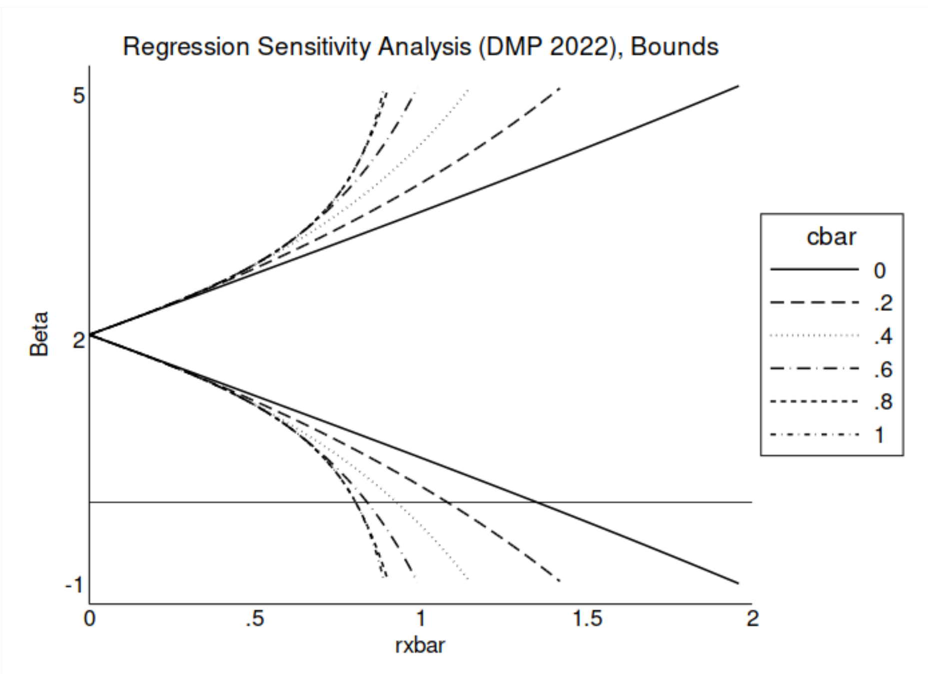
To compare multiple values of $\bar{c}$, a `numlist` can be given in the `cbar` option. With multiple values of $\bar{c}$, the command will show a plot rather than displaying the results in the console.

```
. regsensitivity bounds `y´ `x´ `w´, compare(`w1´) cbar(0(.2)1)
```



By default, the plot will try to show where the identified set becomes $\mathbb{R}$ for each value of $\bar{c}$. For this example, the plot is dominated visually by the identified set for $\bar{c} = 0$. To see better where the identified set instersects with 0, we can rerun the analysis restricting the range of $\bar{r}_X$.

```
. regsensitivity bounds `y´ `x´ `w´, compare(`w1´) cbar(0(.2)1) rxbar(0 2) plot
```



Breakdown Frontier

The output of `regsensitivity bounds` shows a *breakdown point* for a given hypothesis about the parameter β_{long}. For a hypothesis $\beta_{\text{long}} \in B \subseteq \mathbb{R}$, the breakdown point is the smallest value of the sensitivity parameter $\bar{r}_X$ for which the hypothesis does not hold for every β in the identified set. Formally,

$$\bar{r}_X^{bp}(\bar{r}_Y, \bar{c}; B) = \inf\{\bar{r}_X \geq 0 : b \in \mathcal{B}_I(\bar{r}_X, \bar{r}_Y, \bar{c}) \text{ for some } b \in \mathbb{R} \setminus B\}.$$

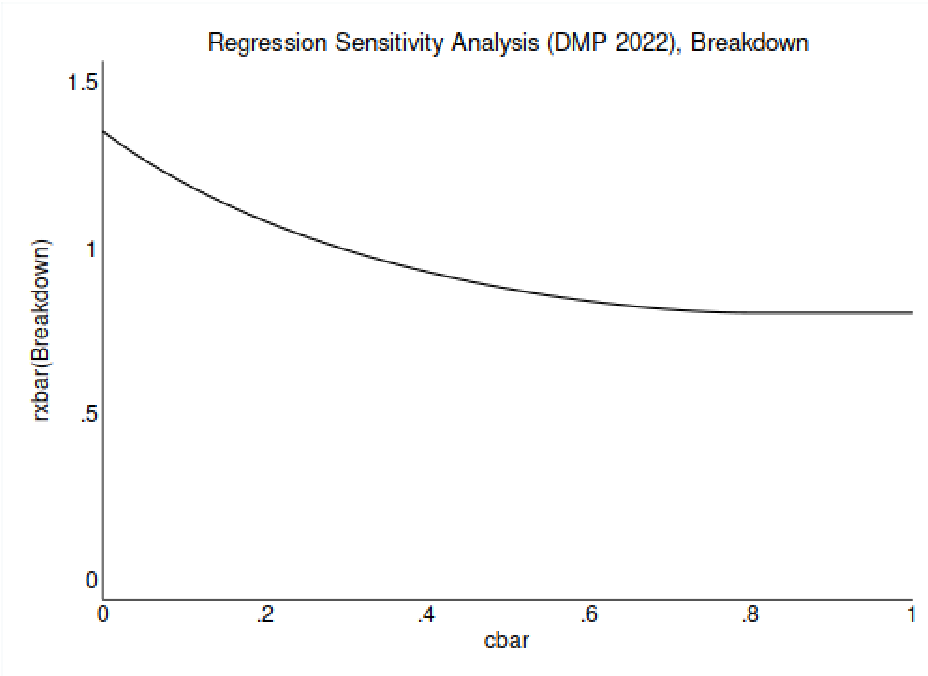
`regsensitivity` can handle hypotheses of the form, $\beta_{\text{long}} \gtrless b$ for any value of b. The default hypothesis is that $\text{sign}(\beta_{\text{long}}) = \text{sign}(\beta_{\text{med}})$, where β_{med} is the coefficient on X in a regression of Y on $(1, X, W_0, W_1)$. In this case $\beta_{\text{med}} > 0$, so the default is to test the hypothesis that $\beta_{\text{long}} > 0$.

The output to `regsensitivity bounds` showed that with W_1 excluding state fixed effects, $\bar{r}_X^{bp}(.1, +\infty; (-\infty, 0]) = 1.195$. To see how this breakdown point varies with the choice of the sensitivity parameter $\bar{c}$, use the `breakdown` subcommand,

```
. regsensitivity breakdown `y´ `x´ `w´, compare(`w1´) cbar(0(.1)1)

Regression Sensitivity Analysis, Breakdown Frontier

Analysis         : DMP (2022)               Number of obs     =      2,036
                                            Beta(short)       =      1.925
Treatment        : tye_tfe890_500kNI_100_l6 Beta(medium)      =      2.055
Outome           : avgrep2000to2016         R2(short)         =      0.033
                                            R2(medium)        =      0.105
                                            Var(Y)            =    101.739
                                            Var(X)            =      0.901
Hypothesis       : Beta > 0                 Var(X_Residual)   =      0.882
Other Params     : rybar = +inf
--------------------------------------------------------------------------

 cbar                             rxbar(Breakdown)
--------------------------------------------------------------------------

 0.000                            135.0%
 0.100                            119.5%
 0.200                            108.0%
 0.300                            99.3 %
 0.400                            92.7 %
 0.500                            87.6 %
 0.600                            83.9 %
 0.700                            81.4 %
 0.800                            80.4 %
 0.900                            80.4 %
 1.000                            80.4 %
--------------------------------------------------------------------------
```

These results can also be plotted using the `plot` subcommand,

```
. regsensitivity plot
```



To test the hypothesis that $\beta_{\text{long}} > b$ for some other value, b, specify `beta(b lb)` (lb for "lower bound"). The `beta` option can also accept a `numlist` to test a range of hypotheses. For example, the following tests the hypotheses that $\beta_{\text{long}} > b$ for a range of values of b,

```
. regsensitivity breakdown `y´ `x´ `w´, compare(`w1´) beta(-1(.2)1 lb)

Regression Sensitivity Analysis, Breakdown Frontier

Analysis         : DMP (2022)                Number of obs     =       2,036
                                             Beta(short)       =       1.925
Treatment        : tye_tfe890_500kNI_100_16  Beta(medium)      =       2.055
Outome           : avgrep2000to2016          R2(short)         =       0.033
                                             R2(medium)        =       0.105
                                             Var(Y)            =     101.739
                                             Var(X)            =       0.901
Hypothesis       : Beta > Beta(Hypothesis)   Var(X_Residual)   =       0.882
Other Params     : cbar = 1, rybar = +inf
----------------------------------------------------------------------------
 Beta(Hypothesis)                   rxbar(Breakdown)
----------------------------------------------------------------------------
 -1.000                             89.1 %
 -0.800                             87.9 %
 -0.600                             86.5 %
 -0.400                             84.8 %
 -0.200                             82.8 %
 0.000                              80.4 %
 0.200                              77.4 %
 0.400                              73.8 %
 0.600                              69.5 %
 0.800                              64.1 %
 1.000                              57.5 %
----------------------------------------------------------------------------
```

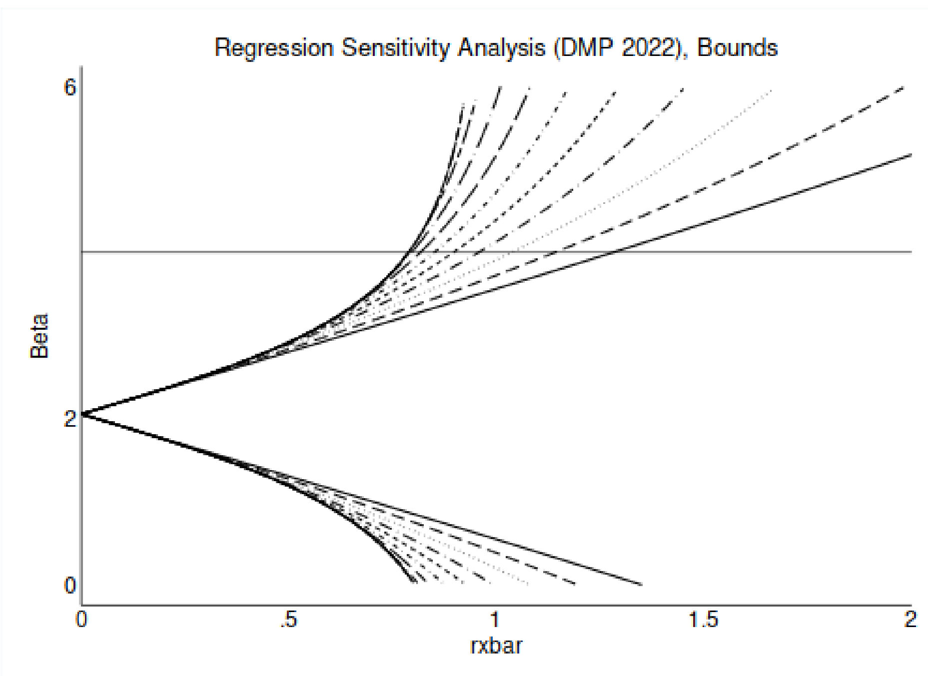
We can also test a hypothesis of the form $\beta < b$, by specifying `beta(b ub)` (ub for "upper bound"). For example the following checks the hypothesis that $\beta < 4$,

```
. regsensitivity breakdown `y´ `x´ `w´, compare(`w1´) cbar(0(.1)1) beta(4 ub)
```

```
Regression Sensitivity Analysis, Breakdown Frontier

Analysis        : DMP (2022)                    Number of obs     =      2,036
                                                Beta(short)       =      1.925
Treatment       : tye_tfe890_500kNI_100_l6      Beta(medium)      =      2.055
Outome          : avgrep2000to2016              R2(short)         =      0.033
                                                R2(medium)        =      0.105
                                                Var(Y)            =    101.739
                                                Var(X)            =      0.901
Hypothesis      : Beta < 4                      Var(X_Residual)   =      0.882
Other Params    : rybar = +inf
-------------------------------------------------------------------------------
 cbar                               rxbar(Breakdown)
-------------------------------------------------------------------------------
 0.000                              128.1%
 0.100                              114.0%
 0.200                              103.6%
 0.300                              95.7 %
 0.400                              89.6 %
 0.500                              85.0 %
 0.600                              81.7 %
 0.700                              79.5 %
 0.800                              78.8 %
 0.900                              78.8 %
 1.000                              78.8 %
-------------------------------------------------------------------------------
```

To see visually where the identified set intersects with 4 we can specify this alternative hypothesis in the **regsensitivity bounds** command. By including the option **beta(4 ub)** the resulting plot will include a horizontal line at 4,

```
. regsensitivity bounds `y´ `x´ `w´, compare(`w1´) rxbar(0 2) cbar(0(.1)1) beta(4 ub)

. regsensitivity plot, nolegend yrange(0 6)
```



Summary statistics

The output of `regsensitivity bounds` and `regsensitivity breakdown` both include a table of summary statistics. These are as follows,

- `Number of observations`
- `Beta(short)`: The coefficient on X in the regression of Y on $(1, X)$
- `Beta(medium)`: The coefficient on X in the regression of Y on $(1, X, W_1)$
- `R2(short)`: The R-squared from the regression of Y on $(1, X)$
- `R2(medium)`: The R-squared from the regression of Y on $(1, X, W_1)$
- `Var(Y)`: Variance of Y
- `Var(X)`: Variance of X
- `Var(X_Residual)`: Variance of $X^{\perp W_1}$, the residual from the regression of X on $(1, W_1)$.

Note: For all the summary statistics reported in this table, (Y, X, W_1) are shorthand for $(Y^{\perp W_0}, X^{\perp W_0}, W_1^{\perp W_0})$ where $Y^{\perp W_0}$ is the residual from the regression of Y on $(1, W_0)$ and likewise for $X^{\perp W_0}$ and $W_1^{\perp W_0}$.

References

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse. 2020. "Frontier Culture: The Roots and Persistence of Rugged Individualism in the United States." *Econometrica* 88 (6): 2329–68. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16484.

Diegert, Paul, Matt Masten, and Alex Poirier. 2022. "Assessing Omitted Variable Bias When the Controls Are Endogenous." *arXiv Preprint*. https://arxiv.org/pdf/2206.02303.pdf.