**DEMO 1: Basic regression modeling in impact evaluation**

/\* Workshop of program impact evaluation. MEASURE Evaluation-INSP, 2015\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/\* Remember that we are trying to estimate the impact of the *Oportunidades* program on the use of contraceptive methods\*/

**/\*Use the dataset with post-information \*/**

\*\*\* use “C:\X\demo1\_CPR.dta”, clear

/\* 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.\*/

. 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)

\*\*\*\*\* Some hypothesis tests using the most simple procedures: chi2 and proportion tests. Remember to restrict the models to those who are eligible.

\*\*\*\* Chi2 test

. tab dich\_contracep program if eligible==1, col chi2

+-------------------+

| Key |

|-------------------|

| frequency |

| column percentage |

+-------------------+

currently |

uses | program or control

contracept | locality

ion? | control program | Total

-----------+----------------------+----------

no | 371 536 | 907

| 82.63 76.14 | 78.66

-----------+----------------------+----------

yes | 78 168 | 246

| 17.37 23.86 | 21.34

-----------+----------------------+----------

Total | 449 704 | 1,153

| 100.00 100.00 | 100.00

Pearson chi2(1) = 6.8837 Pr = 0.009

\*\*\*\* … or we could use a proportion test (by the way, totally equivalent to the chi2 test):

prtest dich\_contracep if eligible==1, by(program )

Two-sample test of proportions control: Number of obs = 449

program: Number of obs = 704

------------------------------------------------------------------------------

Variable | Mean Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

control | .1737194 .0178799 .1386754 .2087633

program | .2386364 .0160649 .2071497 .270123

-------------+----------------------------------------------------------------

diff | -.064917 .0240369 -.1120284 -.0178056

| under Ho: .0247427 -2.62 0.009

------------------------------------------------------------------------------

diff = prop(control) - prop(program) z = -2.6237

Ho: diff = 0

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(Z < z) = 0.0043 Pr(|Z| < |z|) = 0.0087 Pr(Z > z) = 0.9957

\*\*\*\* What is the problem with these estimates? If there is a problem, is it bias? Or is it something else?

\*\*\*\* Let’s use a better approach: A probit model, adjusting for some covariates to increase precision. Due to the fact that units of analysis are clustered at locality (area) level, we have to control for the correlation of the observations within each locality.

. probit dich\_contracep program i.state score age if eligible==1, cluster(locality)

Iteration 0: log pseudolikelihood = -597.68051

Iteration 1: log pseudolikelihood = -551.5078

Iteration 2: log pseudolikelihood = -549.90949

Iteration 3: log pseudolikelihood = -549.90294

Iteration 4: log pseudolikelihood = -549.90294

Probit regression Number of obs = 1153

Wald chi2(8) = 56.16

Prob > chi2 = 0.0000

Log pseudolikelihood = -549.90294 Pseudo R2 = 0.0799

(Std. Err. adjusted for 340 clusters in locality)

--------------------------------------------------------------------------------

| Robust

dich\_contracep | Coef. Std. Err. z P>|z| [95% Conf. Interval]

---------------+----------------------------------------------------------------

program | .2409412 .0996415 2.42 0.016 .0456475 .4362349

|

state |

13 | 1.204251 .2417279 4.98 0.000 .730473 1.678029

21 | .9888095 .2318531 4.26 0.000 .5343858 1.443233

22 | 1.164842 .2931846 3.97 0.000 .5902109 1.739473

24 | 1.424884 .2519225 5.66 0.000 .9311251 1.918643

30 | 1.073736 .2188048 4.91 0.000 .6448866 1.502585

|

score | -.0005395 .0006645 -0.81 0.417 -.0018418 .0007629

age | .1435143 .032097 4.47 0.000 .0806054 .2064232

\_cons | -4.844383 .9678264 -5.01 0.000 -6.741288 -2.947478

--------------------------------------------------------------------------------.

/\*How to interpret the coefficients?\*/

/\*The magnitude of the coefficients by itself is not very meaningful, we need to examine the impact on the prevalence using simulations\*/

g programold=program /\*backup the original variable\*/

. predict p if e(sample) /\*predicted value\*/

(option pr assumed; Pr(dich\_contracep))

(1086 missing values generated)

.

. replace program=0 /\*simulate all without program\*/

(2239 real changes made)

.

. predict p0 if e(sample)

(option pr assumed; Pr(dich\_contracep))

(1086 missing values generated)

.

. replace program=1 /\*simulate all with program\*/

(2239 real changes made)

.

. predict p1 if e(sample)

(option pr assumed; Pr(dich\_contracep))

(1086 missing values generated)

.

. sum p p0 p1 if e(sample)

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

p | 1153 .2134679 .1106089 .0087506 .5199375

p0 | 1153 .1750719 .0916286 .0087506 .4242845

p1 | 1153 .2386416 .1136685 .0163779 .5199375

. /\*The impact is [with program - without program] = p1-p0\*/

.

. g impact=p1-p0

(1086 missing values generated)

sum impact

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

impact | 1153 .0635697 .0229169 .0076273 .095653

. drop program

. rename programold program /\*recovering the original variable\*/

\*\*\*\* We have a second option: if we fit an OLS model, the coefficients are directly interpretable! The cluster option also corrects for heteroskedasticity but if your data is no clustered, you need to specify the “robust” option.

reg dich\_contracep program i.state score age if eligible==1, cluster(locality)

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

|

state |

13 | .2236205 .0457285 4.89 0.000 .133673 .3135679

21 | .1580735 .034607 4.57 0.000 .090002 .2261451

22 | .20816 .0656321 3.17 0.002 .0790627 .3372574

24 | .297262 .0556132 5.35 0.000 .1878716 .4066524

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

------------------------------------------------------------------------------

The program coefficient gives us the impact estimate directly.