

DISCRETIZE: Command to Convert a Continuous Instrument into a Dummy Variable for Instrumental Variable Estimation

Federico Curci, Sébastien Fontenay & Federico Masera

Stata Users Group Meeting, London - Sept. 5-6, 2019

Table of Contents

- 1 Motivations
- 2 `discretize` command
- 3 Illustration

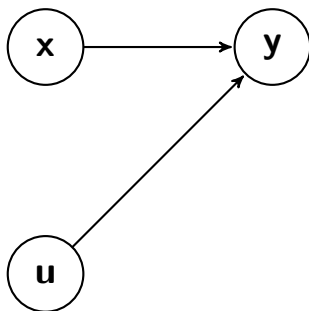
Table of Contents

1 Motivations

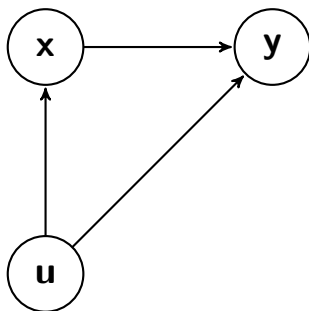
2 discretize command

3 Illustration

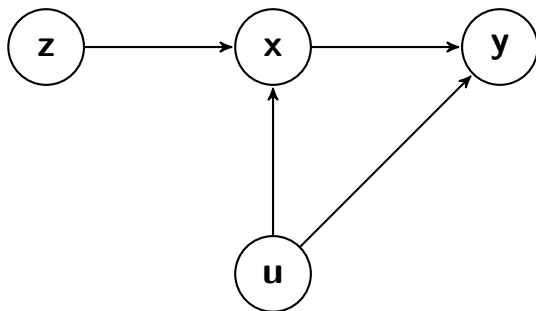
Simple regression model assumes X is uncorrelated with the errors U



If there is an association between X and U : **endogeneity bias**
→ omitted variable, measurement error or simultaneity



Instrumental Variable (IV): instrument Z excluded from outcome equation (second stage), but determinant of endogenous X (first stage)



Motivation

⇒ Researchers often have no *a priori* knowledge or theoretical understanding regarding the relation between Z and X

Motivation

⇒ Researchers often have no *a priori* knowledge or theoretical understanding regarding the relation between Z and X

First Solution: Estimate **(complex) non-linear model** between Z and X

- Advantages:
 - ▶ Strong first stage
 - ▶ Data-driven procedure
- Risks:
 - ▶ Overfitting → including variation of U that we wanted to eliminate

Motivation

⇒ Researchers often have no *a priori* knowledge or theoretical understanding regarding the relation between Z and X

Second Solution: Estimate **linear model** between Z and X

- Advantages:
 - ▶ No overfitting
- Risks:
 - ▶ If relationship is non-linear, often leads to a weak instrument problem

Motivation

⇒ Researchers often have no *a priori* knowledge or theoretical understanding regarding the relation between Z and X

Third Solution: Convert continuous Z into binary instrument

- Advantages:
 - ▶ Parsimonious non-parametric model (Angrist & Pischke, 2009)
 - ▶ Easy interpretation of the variation used
- Risks:
 - ▶ Choice of boundaries for binary instrument often arbitrary
 - ▶ Sensitivity of second stage results

Table of Contents

1 Motivations

2 discretize command

3 Illustration

discretize command

The `discretize` command offers a **data-driven procedure** to build discrete instruments → boundaries chosen to maximize F-statistic in first stage

Main advantages:

- 1 Parsimonious model for first stage relation - No overfitting
- 2 By construction, minimizes weak instrument problem
- 3 Transparent procedure that does not depend on arbitrary decisions made by the researcher
- 4 Graphs to check robustness of second stage results

First stage estimation

```
discretize contvarname, endogenous(varname)  
           range(min/max) interval(min(step)max)
```

contvarname = continuous instrument to be discretized (integer because loops do not handle well decimals)

endogenous(*varname*) = endogenous variable

range(*min/max*) = minimum/maximum values of range

interval(*min(step)max*) = minimum/maximum width of interval

Second stage estimation

```
discretize contvarname, endogenous(varname)  
          range(min/max) interval(min(step)max)  
          second depvar(varname)
```

One needs to specify also `second` and the name of the dependent variable with `depvar(varname)`

Estimation performed using the command `ivregress` with the two-stage least squares (2sls) estimator

Available options

`exogenous(varlist)` exogenous variable(s) used in first and second stage

`interact(varname)` interaction with discretized instrument

`xt(estimator)` panel-data estimators available with the commands `xtreg` and `xtivreg`

`vce(vcetype)` for robust or cluster standard errors

`print` displays values contained in matrix 'results'

`save` saves file with variables stored in matrix 'results' + 95% CI

`graph(string)` graph coefficient estimates (coef) or F-statistics (ftstat)

Table of Contents

1 Motivations

2 discretize command

3 Illustration

Example from Curci & Masera (2018): Rise of violent crime in city centers and suburbanization

Instrument used: **lead poisoning**

- Heavy metal that generates violent behavior
- People exposed principally through car emissions
- Most commonly because lead mixed with soil dust
- Lead is less dangerous when mixed with neutral pH soil

⇒ Chemical theory predicts that cities with neutral soil (around 6.5-7.5 pH) should have lower increase in violent crime

After first stage estimation, the matrix 'results' stores:
Instruments' boundaries, F-statistic, parameter estimate of discrete instrument and standard error

```
. discretize ph10, range(65/80) interval(5(1)10) endogenous(totnpcc_cc_offenses_vc)
> exogenous(i.year) interact(tetra_corr) xt(fe) graph(fstat) print
```

```
results[51,5]
```

	lb	ub	fstat	beta	se
r1	68	77	262.16462	-.00527984	.00032609
r2	68	76	234.77293	-.00515082	.00033617
r3	69	77	227.45227	-.00527996	.00035009
r4	68	78	223.39974	-.00461751	.00030893
r5	68	75	222.05374	-.00523717	.00035145
r6	67	77	207.42131	-.00451308	.00031336
r7	69	76	201.19534	-.0051533	.00036331
r8	70	77	199.14216	-.00526872	.00037336
r9	71	77	199.14216	-.00526872	.00037336
r10	65	75	191.22497	-.00381797	.0002761
r11	69	75	189.88088	-.00529106	.00038397
r12	69	78	188.03554	-.00449492	.00032779
r13	67	76	182.06497	-.00434235	.00032182
r14	66	76	176.64343	-.00396422	.00029827
r15	72	77	175.57532	-.00550638	.00041556
r16	71	76	173.76344	-.00514243	.00039011
r17	70	76	173.76344	-.00514243	.00039011
r18	68	74	173.53996	-.00487553	.0003701
r19	67	75	168.13245	-.00433725	.00033449
r20	70	75	163.5051	-.00533389	.00041714

We can use the new discrete instrument with boundaries 6.8 and 7.7 that has been found to maximize the F-stat in the first stage

```
. gen good_soil = (phi_plc_wtm_wtm_0_r>=6.8 & phi_plc_wtm_wtm_0_r<=7.7)
. xtivreg perc_cc i.year (standardized_vc = c.good_soil#c.tetra_corr), fe
Fixed-effects (within) IV regression      Number of obs   =    9,481
Group variable: fipsplace_00             Number of groups =     305
R-sq:                                     Obs per group:
    within = .                               min =           8
    between = 0.0855                         avg =          31.1
    overall = 0.0795                         max =           32
                                           Wald chi2(32)   =  633103.54
                                           Prob > chi2     =    0.0000
corr(u_i, Xb) = 0.0259                    Wald chi2(32)   =  633103.54
                                           Prob > chi2     =    0.0000
```

perc_cc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
standardized_vc	-.0717297	.00594	-12.08	0.000	-.0833718	-.0600876
year						
1961	.0017654	.0040017	0.44	0.659	-.0060779	.0096087
...						
1991	.0768294	.0113749	6.75	0.000	.0545349	.0991238
_cons	.4348947	.0031643	137.44	0.000	.4286929	.4410965
sigma_u	.18215015					
sigma_e	.04846004					
rho	.93389896	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(304,9144) =  435.91      Prob > F      =  0.0000
```

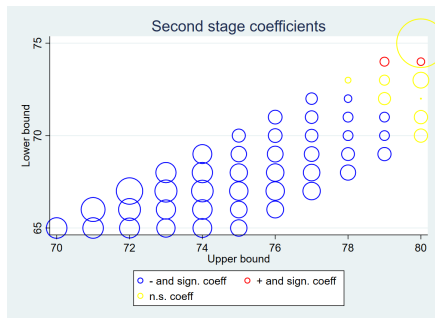
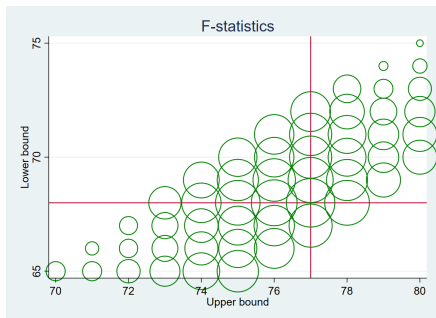
```
Instrumented:  standardized_vc
```

After second stage estimation, the matrix 'results' stores:
Instruments' boundaries, parameter estimate of endogenous variable and standard error

```
. discretize ph10, range(65/80) interval(5(1)10) endogenous(standardized_vc) second
> depvar(perc_cc) exogenous(i.year) interact(tetra_corr) xt(fe) graph(coef) print
results[51,4]
```

	lb	ub	beta	se
r1	70	77	-.04097976	.00580547
r2	71	77	-.04097976	.00580547
r3	69	77	-.05647729	.00583521
r4	68	77	-.07172966	.00593996
r5	68	78	-.05994759	.00599139
r6	69	78	-.042527	.00599988
r7	72	77	-.03381604	.00603609
r8	71	78	-.02463927	.00619798
r9	70	78	-.02463927	.00619798
r10	71	76	-.04882763	.00641164
r11	70	76	-.04882763	.00641164
r12	70	75	-.04405828	.00647297
r13	69	76	-.06484251	.00648862
r14	69	75	-.06214748	.00657464
r15	68	76	-.08023395	.00660769
r16	68	75	-.07907977	.00674165
r17	72	78	-.01415127	.00674563
r18	65	75	-.07021066	.00684718
r19	71	80	-.01309482	.00686332
r20	70	80	-.01309482	.00686332

Graphics allow users to check the sensitivity of the results to the choice of instruments



Package available from:

<https://sites.google.com/view/sebastien-fontenay/stata>

Comments and suggestions are very much welcomed!

sebastien.fontenay@uclouvain.be