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

Federico Curci ¹, Sébastien Fontenay ² & Federico Masera³

¹ Colegio Universitario de Estudios Financieros ² Université Catolique de Louvain ³ University of New South Wales

Oceania Stata Meeting, Parramatta - Aug. 19-20, 2019
Table of Contents

1 Motivations

2 discretiz command

3 Illustration
Table of Contents

1 Motivations

2 discretiz 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**

$\rightarrow$ omitted variable, measurement error or simultaneity
Instrumental Variable (IV): instrument Z excluded from outcome equation (second stage), but determinant of endogenous X (first stage)
Motivations

Researchers often have no \textit{a priori} knowledge or theoretical understanding regarding the relation between $Z$ and $X$ which can lead to \textbf{model misspecification}.
Motivations

Researchers often have no *a priori* knowledge or theoretical understanding regarding the relation between $Z$ and $X$ which can lead to model misspecification.

If the model is in fact non-linear, fitting a linear model for the first stage could lead to a problem of *weak instrument*.
Motivations

Researchers often have no *a priori* knowledge or theoretical understanding regarding the relation between Z and X which can lead to model misspecification.

If the model is in fact non-linear, fitting a linear model for the first stage could lead to a problem of *weak instrument*.

Solution proposed by Angrist & Pischke (2009) to convert continuous Z into binary instrument which provides parsimonious non-parametric model for the underlying first stage relation.
Motivations

Researchers often have no \textit{a priori} knowledge or theoretical understanding regarding the relation between $Z$ and $X$ which can lead to \textbf{model misspecification}

If the model is in fact non-linear, fitting a linear model for the first stage could lead to a problem of \textbf{weak instrument}

Solution proposed by Angrist & Pischke (2009) to convert continuous $Z$ into binary instrument which provides \textbf{parsimonious non-parametric model} for the underlying first stage relation

Unfortunately, construction of binary instrument \textbf{often appears to be arbitrary}, which may raise concerns about the robustness of the second stage results
Table of Contents

1 Motivations

2 discretiz command

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

Main advantages:

1. Minimizes weak instrument problem that can arise in case of incorrect functional specification in the first stage
2. Transparent procedure that does not depend on arbitrary decisions made by the researcher
First stage estimation

```
discretiz 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

\begin{verbatim}
  discretiz contvarname, endogenous(varname)

  range(min/max) interval(min(step)max)

  second depvar(varname)
\end{verbatim}

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 discretiz command
3 Illustration
Understand if violent crime in city centers affects the spread of cities in the US (movement of people from city centers to suburbs)

Idea for instrument:
- Lead heavy metal that in case of poisoning generates violent behavior
- People are exposed to lead through car emissions
- Most common method of contact: lead mixed with soil dust
- Lead is less dangerous when mixed with neutral pH soil

Time variation: After the end of WW2 lead poisoning increase dramatically. Decreased after 1972 because of lead use regulation

Cross-sectional variation: pH of the soil of different cities

Chemical theory predicts that during the high lead use years cities with neutral soil (around the 6.5-7.5 pH) should have less of an increase in violent crime.
After first stage estimation, the matrix ‘results’ stores: Instruments’ boundaries, F-statistic, parameter estimate of discrete instrument and standard error

```
.discretiz 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]

<table>
<thead>
<tr>
<th></th>
<th>lb</th>
<th>ub</th>
<th>fstat</th>
<th>beta</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>68</td>
<td>77</td>
<td>262.16462</td>
<td>-.00527984</td>
<td>.00032609</td>
</tr>
<tr>
<td>r2</td>
<td>68</td>
<td>76</td>
<td>234.77293</td>
<td>-.00515082</td>
<td>.00033617</td>
</tr>
<tr>
<td>r3</td>
<td>69</td>
<td>77</td>
<td>227.45227</td>
<td>-.00527996</td>
<td>.00035009</td>
</tr>
<tr>
<td>r4</td>
<td>68</td>
<td>78</td>
<td>223.39974</td>
<td>-.00461751</td>
<td>.00030893</td>
</tr>
<tr>
<td>r5</td>
<td>68</td>
<td>75</td>
<td>222.05374</td>
<td>-.00523717</td>
<td>.00035145</td>
</tr>
<tr>
<td>r6</td>
<td>67</td>
<td>77</td>
<td>207.42131</td>
<td>-.00451308</td>
<td>.00031336</td>
</tr>
<tr>
<td>r7</td>
<td>69</td>
<td>76</td>
<td>201.19534</td>
<td>-.0051533</td>
<td>.00036331</td>
</tr>
<tr>
<td>r8</td>
<td>70</td>
<td>77</td>
<td>199.14216</td>
<td>-.00526872</td>
<td>.00037336</td>
</tr>
<tr>
<td>r9</td>
<td>71</td>
<td>77</td>
<td>199.14216</td>
<td>-.00526872</td>
<td>.00037336</td>
</tr>
<tr>
<td>r10</td>
<td>65</td>
<td>75</td>
<td>191.22497</td>
<td>-.00381797</td>
<td>.0002761</td>
</tr>
<tr>
<td>r11</td>
<td>69</td>
<td>75</td>
<td>189.88088</td>
<td>-.00529106</td>
<td>.00038397</td>
</tr>
<tr>
<td>r12</td>
<td>69</td>
<td>78</td>
<td>188.03554</td>
<td>-.00449492</td>
<td>.00032779</td>
</tr>
<tr>
<td>r13</td>
<td>67</td>
<td>76</td>
<td>182.06497</td>
<td>-.00434235</td>
<td>.00032182</td>
</tr>
<tr>
<td>r14</td>
<td>66</td>
<td>76</td>
<td>176.64343</td>
<td>-.00396422</td>
<td>.00029827</td>
</tr>
<tr>
<td>r15</td>
<td>72</td>
<td>77</td>
<td>175.57532</td>
<td>-.00550638</td>
<td>.00041556</td>
</tr>
<tr>
<td>r16</td>
<td>71</td>
<td>76</td>
<td>173.76344</td>
<td>-.00514243</td>
<td>.00039011</td>
</tr>
<tr>
<td>r17</td>
<td>70</td>
<td>76</td>
<td>173.76344</td>
<td>-.00514243</td>
<td>.00039011</td>
</tr>
<tr>
<td>r18</td>
<td>68</td>
<td>74</td>
<td>173.53996</td>
<td>-.00487553</td>
<td>.0003701</td>
</tr>
<tr>
<td>r19</td>
<td>67</td>
<td>75</td>
<td>168.13245</td>
<td>-.00433725</td>
<td>.00033449</td>
</tr>
<tr>
<td>r20</td>
<td>70</td>
<td>75</td>
<td>163.5051</td>
<td>-.00533389</td>
<td>.00041714</td>
</tr>
</tbody>
</table>
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 = (ph1_plc_wtm_wtm_0_r>=6.8 & ph1_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*

| Coef. Std. Err. | z    | P>|z|  | [95% Conf. Interval] |
|------------------|------|------|----------------------|
| perc_cc          |      |      |                      |
| standardized_vc  | -0.0717297 | 0.00594 | -12.08 | 0.000 | -0.0833718, -0.0600876 |
| year             |      |      |                      |
| 1961             | 0.0017654 | 0.0040017 | 0.44 | 0.659 | -0.0060779, 0.0096087 |
| ...              |      |      |                      |
| 1991             | 0.0768294 | 0.0113749 | 6.75 | 0.000 | 0.0545349, 0.0991238 |
| _cons            | 0.4348947 | 0.0031643 | 137.44 | 0.000 | 0.4286929, 0.4410965 |

| sigma_u          | 0.18215015 |
| sigma_e          | 0.04846004 |
| rho              | 0.93389896 (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

```
  discretiz 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  -0.04097976  0.00580547
     r2  71  77  -0.04097976  0.00580547
     r3  69  77  -0.05647729  0.00583521
       r4  68  77  -0.07172966  0.00593996
     r5  68  78  -0.05994759  0.00599139
     r6  69  78  -0.042527     0.00599888
     r7  72  77  -0.03381604  0.00603609
     r8  71  78  -0.02463927  0.00619798
     r9  70  78  -0.02463927  0.00619798
    r10  71  76  -0.04882763  0.00641164
    r11  70  76  -0.04882763  0.00641164
    r12  70  75  -0.04405828  0.00647297
    r13  69  76  -0.06484251  0.00648862
    r14  69  75  -0.06214748  0.00657464
    r15  68  76  -0.08023395  0.00660769
    r16  68  75  -0.07907977  0.00674165
    r17  72  78  -0.01415127  0.00674563
    r18  65  75  -0.07021066  0.00684718
    r19  71  80  -0.01309482  0.00686332
    r20  70  80  -0.01309482  0.00686332
```
Graphics allow users to check the sensitivity of the results to the choice of instruments.