Cross-fit partial out problem set

Marek Chadim

1 October 2024

```
. clear all

. use penn_jae

. keep if tg ==0 | tg == 4
(8,814 observations deleted)

. rename tg D

. replace D = 1 if D == 4
(1,745 real changes made)

. gen Y = ln(inuidur1)
```

```
. foreach x of varlist female-husd{
  2. foreach y of varlist female-husd {
  3. generate `x'X`y'=`x'*`y'
  4. }
  5. }
```

```
. vl set

-------------------------------------------------------------------------------------
                     |                      Macro's contents
                     |------------------------------------------------------------------
Macro                |  # Vars   Description
---------------------+------------------------------------------------------------------
System               |
  $vlcategorical     |     323   categorical variables
  $vlcontinuous      |       1   continuous variable
  $vluncertain       |       3   perhaps continuous, perhaps categorical variables
  $vlother           |      61   all missing or constant variables
-------------------------------------------------------------------------------------
Notes

      1. Review contents of vlcategorical and vlcontinuous to ensure they are c

      2. If there are any variables in vluncertain, you can reallocate them to 

      3. Use vl move to move variables among classifications.  For example, typ
         classification.

      4. vlnames are global macros.  Type the vlname without the leading dollar
         dollar sign with other Stata commands to get a varlist.

. vl move (D) vlother
note: 1 variable specified and 1 variable moved.

------------------------------
Macro          # Added/Removed
------------------------------
$vlcategorical              -1
$vlcontinuous                0
$vluncertain                 0
$vlother                     1
------------------------------

. vl substitute ifactors = i.vlcategorical

. display "$ifactors"
i.female i.black i.hispanic i.othrace i.dep i.q1 i.q2 i.q3 i.q4 i.v14 i.q5 i.q6
> eXblack i.femaleXhispanic i.femaleXothrace i.femaleXdep i.femaleXq1 i.femaleX
> gelt35 i.femaleXagegt54 i.femaleXdurable i.femaleXnondurable i.femaleXlusd i.
> ackXv14 i.blackXq5 i.blackXq6 i.blackXrecall i.blackXagelt35 i.blackXagegt54 
> hispanic i.hispanicXdep i.hispanicXq2 i.hispanicXq3 i.hispanicXq4 i.hispanicX
> hispanicXdurable i.hispanicXnondurable i.hispanicXlusd i.hispanicXhusd i.othr
> eXv14 i.othraceXq5 i.othraceXrecall i.othraceXagelt35 i.othraceXagegt54 i.oth
> .depXhispanic i.depXothrace i.depXdep i.depXq1 i.depXq2 i.depXq3 i.depXq4 i.d
```

```
> urable i.depXlusd i.depXhusd i.q1Xfemale i.q1Xdep i.q1Xq1 i.q1Xq6 i.q1Xrecall
> 2Xothrace i.q2Xdep i.q2Xq2 i.q2Xq6 i.q2Xrecall i.q2Xagelt35 i.q2Xagegt54 i.q2
> race i.q3Xdep i.q3Xq3 i.q3Xq6 i.q3Xrecall i.q3Xagelt35 i.q3Xagegt54 i.q3Xdura
> i.q4Xdep i.q4Xq4 i.q4Xq6 i.q4Xrecall i.q4Xagelt35 i.q4Xagegt54 i.q4Xdurable i
> .v14Xdep i.v14Xv14 i.v14Xq6 i.v14Xrecall i.v14Xagelt35 i.v14Xagegt54 i.v14Xdu
> race i.q5Xdep i.q5Xq5 i.q5Xq6 i.q5Xrecall i.q5Xagelt35 i.q5Xagegt54 i.q5Xdura
> Xq1 i.q6Xq2 i.q6Xq3 i.q6Xq4 i.q6Xv14 i.q6Xq5 i.q6Xq6 i.q6Xrecall i.q6Xagelt35
> ack i.recallXhispanic i.recallXothrace i.recallXdep i.recallXq1 i.recallXq2 i
> 54 i.recallXdurable i.recallXnondurable i.recallXlusd i.recallXhusd i.agelt35
> agelt35Xq2 i.agelt35Xq3 i.agelt35Xq4 i.agelt35Xv14 i.agelt35Xq5 i.agelt35Xq6
> i.agelt35Xhusd i.agegt54Xfemale i.agegt54Xblack i.agegt54Xhispanic i.agegt54X
> agegt54Xq5 i.agegt54Xq6 i.agegt54Xrecall i.agegt54Xagelt35 i.agegt54Xagegt54
> bleXhispanic i.durableXothrace i.durableXdep i.durableXq1 i.durableXq2 i.dura
> lt35 i.durableXdurable i.durableXnondurable i.durableXlusd i.durableXhusd i.n
> eXdep i.nondurableXq2 i.nondurableXq3 i.nondurableXq4 i.nondurableXv14 i.nond
> nondurableXdurable i.nondurableXnondurable i.lusdXfemale i.lusdXblack i.lusdX
> sdXq5 i.lusdXq6 i.lusdXrecall i.lusdXagelt35 i.lusdXagegt54 i.lusdXdurable i.
> .husdXq3 i.husdXq4 i.husdXv14 i.husdXq5 i.husdXq6 i.husdXrecall i.husdXagelt3
```

Use the cross-fit partialing out estimator and set the tuning parameter using the plugin formula

```
. qui: xporegress Y D, controls($ifactors) xfolds(5) resample(15) rseed(42) sel

. etable

------------------------------
                             Y
------------------------------
Treatment group          -0.086
                        (0.037)
Number of observations     5099
------------------------------
```

The results are similar to Table 1, column 1 of Chernozhukov et al (2018).

The selected variables for a specific fold and sample are

```
. dis e(k_controls_sel)
33

. dis e(controls_sel)
0bn.agegt54Xq1 0bn.agelt35Xhusd 0bn.agelt35Xq1 0bn.black 0bn.blackXblack 0bn.bl
> agelt35 0bn.q2Xq6 0bn.q6 0bn.q6Xq2 0bn.q6Xq6 0bn.q6Xrecall 0bn.recallXblack 0
> black 1bn.nondurableXrecall 1bn.q1Xagelt35 1bn.q1Xdep 1bn.q2Xq6 1bn.q6 1bn.q6
```

Use the same specification but now select the tuning parameter by cross-validation

```
. qui: xporegress Y D, controls($ifactors) xfolds(5) resample(15) rseed(42) sel

. etable

------------------------------
                             Y
------------------------------
Treatment group          -0.085
                        (0.036)
Number of observations     5099
------------------------------
```

The number of variables included is

```
. dis e(k_controls_sel)
114
```

This number higher because cross-validation typically underpenalizes to reduce bias to obtain better estimates but tends to select a substantial larger number of variables. This suggests cross-validation to be more suitable for the algorithm based on optimal instrument than for the algorithm based on double selection (Belloni, Chernozhukov & Wei, 2016).

```
. vl move (female-husd) vlother
note: 19 variables specified and 19 variables moved.

------------------------------
Macro          # Added/Removed
------------------------------
$vlcategorical             -19
$vlcontinuous                0
$vluncertain                 0
$vlother                    19
------------------------------

. vl substitute ifactors2 = i.vlcategorical

. display "$ifactors2"
i.femaleXfemale i.femaleXblack i.femaleXhispanic i.femaleXothrace i.femaleXdep
> aleXrecall i.femaleXagelt35 i.femaleXagegt54 i.femaleXdurable i.femaleXnondur
> ckXq3 i.blackXq4 i.blackXv14 i.blackXq5 i.blackXq6 i.blackXrecall i.blackXage
> icXfemale i.hispanicXhispanic i.hispanicXdep i.hispanicXq2 i.hispanicXq3 i.his
> i.hispanicXagegt54 i.hispanicXdurable i.hispanicXnondurable i.hispanicXlusd i
> i.othraceXq4 i.othraceXv14 i.othraceXq5 i.othraceXrecall i.othraceXagelt35 i.
> XfemaleXblack i.depXhispanic i.depXothrace i.depXdep i.depXq1 i.depXq2
> epXdurable i.depXnondurable i.depXlusd i.depXhusd i.q1Xfemale i.q1Xdep i.q1Xq
> ack i.q2Xhispanic i.q2Xothrace i.q2Xdep i.q2Xq2 i.q2Xq6 i.q2Xrecall i.q2Xagel
> .q3Xhispanic i.q3Xothrace i.q3Xdep i.q3Xq3 i.q3Xq6 i.q3Xrecall i.q3Xagelt35 i
> ispanic i.q4Xothrace i.q4Xdep i.q4Xq4 i.q4Xq6 i.q4Xrecall i.q4Xagelt35 i.q4Xa
> panic i.v14Xothrace i.v14Xdep i.v14Xv14 i.v14Xq6 i.v14Xrecall i.v14Xagelt35 i
> .q5Xhispanic i.q5Xothrace i.q5Xdep i.q5Xq5 i.q5Xq6 i.q5Xrecall i.q5Xagelt35 i
> ispanic i.q6Xdep i.q6Xq1 i.q6Xq2 i.q6Xq3 i.q6Xq4 i.q6Xv14 i.q6Xq5 i.q6Xq6 i.q
> llXfemale i.recallXblack i.recallXhispanic i.recallXothrace i.recallXdep i.re
> recall i.recallXagegt54 i.recallXdurable i.recallXnondurable i.recallXlusd i.
> 5Xdep i.agelt35Xq1 i.agelt35Xq2 i.agelt35Xq3 i.agelt35Xq4 i.agelt35Xv14 i.age
> rable i.agelt35Xlusd i.agelt35Xhusd i.agegt54Xfemale i.agegt54Xblack i.agegt5
> 4Xq4 i.agegt54Xv14 i.agegt54Xq5 i.agegt54Xq6 i.agegt54Xrecall i.agegt54Xagelt
> .durableXblack i.durableXhispanic i.durableXothrace i.durableXdep i.durableXq
> XrecallXagelt35 i.durableXdurable i.durableXnondurable i.durableXlu
> eXothrace i.nondurableXdep i.nondurableXq2 i.nondurableXq3 i.nondurableXq4 i.
> nondurableXagegt54 i.nondurableXdurable i.nondurableXnondurable i.lusdXfemale
> sdXq4 i.lusdXv14 i.lusdXq5 i.lusdXq6 i.lusdXrecall i.lusdXagelt35 i.lusdXageg
> .husdXdep i.husdXq2 i.husdXq3 i.husdXq4 i.husdXv14 i.husdXq5 i.husdXq6 i.husd
```

Set the raw covariates so that they are always included in the regression and use Lasso to select additional controls among the second-order terms.

```
. qui: xporegress Y D , controls( (female-husd) $ifactors2) xfolds(5) resample(

. etable

------------------------------
                             Y
------------------------------
Treatment group         -0.075
                       (0.036)
Number of observations    5099
------------------------------
```

The number of selected control variables is,

```
. dis e(k_controls_sel)
34
```

and they are

```
. dis e(controls_sel)
0bn.agelt35Xq1 0bn.hispanicXlusd 0bn.lusdXhispanic 0bn.nondurableXdurable 0bn.q
> eXdep 1bn.q1Xagegt54 1bn.q1Xagelt35 1bn.q1Xdep 2bn.q1Xdep agegt54 agelt35 bla
```