DS-HECK: Double-Lasso Estimation of Heckman Selection Model

Di Liu

Principal Econometrician

Stata

Nov 1, 2024

High-dimensional sample selection model

$$y_1 = \mathbf{x}_1'\alpha + u_1 \qquad \text{(main equation)}$$
$$y_2 = \mathbb{I}(\mathbf{x}'\beta + \mathbf{z}'\eta + u_2 \geq 0) \qquad \text{(selection equation)}$$

- y_1 is the outcome of interest
- $\mathbf{x}_1$ is a low dimensional vector of independent variables
- y_2 is a sample selection indicator
- $\mathbf{x}_2 = (\mathbf{x}_1, \mathbf{x}_2)$ is a low dimensional vector
- $\mathbf{z}$ is a high-dimensional vector
- η is a sparse vector

Sparse model for observed outcome

Under some assumptions, the conditional mean of the observed outcome is

$$\begin{aligned}\mathbb{E}(y_1|\mathbf{x},\mathbf{z},y_2=1) &= \mathbf{x}_1'\alpha + \gamma\lambda(\mathbf{x}'\beta + \mathbf{z}'\eta) \\ &= \mathbf{x}_1'\alpha + \gamma\lambda(\mathbf{x}'\beta) + \gamma\lambda^{(1)}(q)\mathbf{z}'\eta \\ &= \mathbf{x}_1'\alpha + \gamma\lambda(\mathbf{x}'\beta) + \mathbf{z}'\omega\end{aligned}$$

Sparse $\eta \implies$ the same sparsity pattern in ω.

Objective

Consistently estimate α and γ with high-dimensional $\mathbf{z}$ with sparse coefficients ω.

- α estimates effects of $\mathbf{x}_1$ on y_1.
- γ estimates the extent of sample selection bias
- ω is nuisance parameter

DS-HECK: Two double-Lassos

Two high-dimensional models:

$$\mathbb{E}(y_1|\mathbf{x},\mathbf{z},y_2=1)=\mathbf{x}_1'\alpha+\gamma\lambda(\mathbf{x}'\beta)+\mathbf{z}'\omega \qquad \text{(Observed outcome)}$$

$$y_2=\mathbb{I}(\mathbf{x}'\beta+\mathbf{z}'\eta+u_2\geq 0) \qquad \text{(selection)}$$

1. If we know β, we can estimate α and γ by running the double-Lasso to the linear regression in Eq. (observed outcome).
2. However, we can consistently estimate β by running the double-lasso to the Probit regression in Eq. (selection).
3. Standand errors must be adjusted because β is estimated.

dsheckman: Stata command for DS-HECK

Syntax

```
dsheckman depvar indepvars [if] [in]
          , selection(depvar_s = indepvars_s)
          [ selvars(varlist) ]
```

Model

$$
\begin{aligned}
y_1 &= \mathbf{x}_1{}'\alpha + u_1 && \text{(main equation)}\\
y_2 &= \mathbb{I}(\mathbf{x}'\beta + \mathbf{z}'\eta + u_2 \geq 0) && \text{(selection equation)}
\end{aligned}
$$

- *depvar* $\equiv y_1$, *indepvars* $\equiv \mathbf{x}_1$
- *depvar_s* $\equiv y_2$, *indepvars_s* $\equiv (\mathbf{x}, \mathbf{z})$
- selvars() $\equiv \mathbf{x}$ if specified. Otherwise, $\mathbf{x}$ is chosen by Lasso.

Example: Labor participation and earnings

$$\log(\texttt{income}) = \alpha_0 + \alpha_1 \cdot \texttt{educ} + \alpha_2 \cdot \texttt{exper} + u_1 \qquad \text{(earning)}$$

$$\texttt{inlf} = \mathbb{I}(\mathbf{x}'\beta + \mathbf{z}'\eta + u_2 \geq 0) \qquad \text{(labor participation)}$$

Step 1: Define $(\mathbf{x}, \mathbf{z})$ in the labor participation equation

```
. local vars_sel exper exper2 educ_level  childcare_expen_2012          ///
>         i.if_kidsle15 num_kids wage_husband exp_appl i.wtr_enrolled    ///
>         i.wtr_grad_hs i.wtr_attend_college i.wtr_cert_educ             ///
>         i.wtr_educ_usa  i.father_educ_usa i.mother_educ_usa            ///
>         i.rural_urban i.own_vehicle i.current_state
```

```
. dsheckman lnwage educ_level exper, selection(inlf = `vars_sel´)

step 1: lasso probit to select vars
step 2: dsprobit of y2 on selected zvars

Double selection probit                 Number of obs              =      1,989
                                        Number of controls         =         89
                                        Number of selected controls =        10

------------------------------------------------------------------------------
             |               Robust
        inlf | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
  educ_level |     .071556   .0228702     3.13   0.002     .0267313    .1163807
      exper2 |   -.0011333   .0003511    -3.23   0.001    -.0018214   -.0004451
childca~2012 |    .0726106   .0258034     2.81   0.005     .0220368    .1231845
       exper |    .0156069   .0119473     1.31   0.191    -.0078093    .0390231
       _cons |   -.7051439   .2897608    -2.43   0.015    -1.273065   -.1372231
------------------------------------------------------------------------------

step 3: compute lambda
step 4: dsregress y1 on xvars, lambda with controls

Double-selection-lasso Heckman          Number of obs              =      1,989
                                               Selected            =      1,294
                                               Nonselected         =        695
                                        Number of variables        =         93
                                        Number of selected controls =         3
                                        Number of main variables   =          2

------------------------------------------------------------------------------
             | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
  educ_level |    .0544304   .0382198     1.42   0.154    -.0204791    .1293399
       exper |    .0320553   .0079836     4.02   0.000     .0164076    .0477029
      lambda |    -1.93624   .4908842    -3.94   0.000    -2.898355   -.9741244
------------------------------------------------------------------------------

Note: in the main equation, there are 2 variables; in the selection equation,
      3 among 93 variables are used to predict inverse mills ratio.
```

option `selvars()`

```
. dsheckman lnwage educ_level exper, selection(inlf = `vars_sel´) ///
>         selvars(num_kids educ_level exper)
```

step 1: set *varsofinterest* in selection equation
step 2: dsprobit of y2 on selected zvars

```
Double selection probit               Number of obs                =      1,989
                                      Number of controls           =         90
                                      Number of selected controls  =         12
```

inlf	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
num_kids	-.1008661	.0440309	-2.29	0.022	-.1871651	-.0145671
educ_level	.0719182	.0229172	3.14	0.002	.0270014	.116835
exper	.0172596	.011989	1.44	0.150	-.0062385	.0407576
_cons	-.7082983	.2901592	-2.44	0.015	-1.277	-.1395968

step 3: compute lambda
step 4: dsregress y1 on xvars, lambda with controls

```
Double-selection-lasso Heckman        Number of obs                =      1,989
                                            Selected               =      1,294
                                            Nonselected            =        695
                                      Number of variables          =         93
                                      Number of selected controls  =          3
                                      Number of main variables     =          2
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
educ_level	.0954287	.0356875	2.67	0.007	.0254825	.1653749
exper	.0155112	.0163068	0.95	0.341	-.0164495	.0474719
lambda	-.9183578	.7577044	-1.21	0.226	-2.403431	.5667155

Note: in the main equation, there are 2 variables; in the selection equation,
3 among 93 variables are used to predict inverse mills ratio.

Resources

https://github.com/flyingliudi/dsheck_public