Learning About Selection: An Improved Correction Procedure

Iain G. Snoddy

27 July 2018

Ph.D. Candidate
Vancouver School of Economics

2018 Canadian Stata Conference

Motivation: Old Method, New Techniques

Question: How to estimate the returns to schooling when people select across locations?

Influential Paper in Economics to control for self-selection: Dahl (2002), *Econometrica*

Dahl's Contribution

- Reduces dimension of problem
- Non-parametric implementation
- Control function approach

Earnings Equation:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + u_{ic}, \quad c = 1, \ldots, C$$

Utility Equation:

$$V_{ijc} = y_{ic} + \pi_{ijc}, \quad c = 1, \ldots, C$$

where $\pi_{ijc} = \gamma_{jc} z_i + \epsilon_{ijc}, \quad c = 1, \ldots, C$

i indexes individuals, c states, j birth state

We can re-write the utility function as:

$$V_{ijc} = \mathbb{E}\left[y_{ic}|s_i, x_i\right] + \mathbb{E}\left[\pi_{ijc}|z_i\right] + \epsilon_{ijc} + u_{ic} = \vartheta_{jc} + \omega_{ijc}$$

The selection rule:

$$y_{ic} \text{ observed} \quad \Longleftrightarrow \quad \max_k \left(\vartheta_{jk} - \vartheta_{jc} + \omega_{ijk} - \omega_{ijc}\right) \leq 0$$

Selection bias:

$$E[u_{ic}|y_{ic} \text{ observed}] = E[u_{ic}|\vartheta_{jc} - \vartheta_{jk} \geq \omega_{ijk} - \omega_{ijc}, \forall k \neq c] \neq 0$$

Full set of migration probabilities summarise the selection problem: $(p_{ij1}, ..., p_{ijN})$

Estimating equation:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + \sum_j M_{ijc} \times \mu_{jc} \left(p_{ij1}, ..., p_{ijN}\right) + v_{ic}$$

Dahl's Assumption

Dahl makes the Single Index Sufficiency Assumption (SISA).

All of the information in $(p_{ij1}, ..., p_{ijN})$ is summarised by p_{ijc}.

Which implies:

$$cov(u_{ic}, \omega_{ijm} - \omega_{ijc}) = K, \quad \forall m \neq k$$

Dahl's Implementation

Estimating Equation:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + \sum_j M_{ijc} \times \hat{\mu}_{jc} \left(p_{ijc} \right) + v_{ic}$$

- Migration probabilities estimated by grouping individuals into cells
- `selmlog13` Stata command by François Bourguignon, Martin Fournier, and Marc Gurgand

Improvement 1: Better P Estimates

- Cell approach involves ad hoc choices
- Alternative: use a Neural Network, or Random Forest
- Ties researchers' hands
- Reduces variance
- Reduces noise from poor predictors

Improvement 2: Better Variable Selection

The SISA is restrictive!

Start with full model:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + \tilde{\mu}_c \left(\hat{p}_{i1}, ..., \hat{p}_{iN} \right) + \tilde{v}_{ic}$$

Use Double-Post LASSO to select included terms!

Improvement 2: Double-Post LASSO

Belloni, Chernozhukov, and Hansen (2014)

LASSO:

$$\min_{\beta} (y - X\beta)^T (y - X\beta) \quad \text{subject to} \quad ||\beta||_1 \leq t$$

where t is a free parameter that determines regularization.

Procedure:

1. Run LASSO of y on terms
2. Run LASSO of x on terms
3. Run y on x plus terms included in 1 & 2

Improvement 2: Does it Work???

Monte Carlo experiment: Use the Roy Model

The SISA: $u_{ic} = \tau_c a_i + b_{ic}$

Three cases:

- SISA holds
- SISA weak violation
- SISA strong violation

Lassopack

Implemented using `Lassopack`- Ahrens, Hansen, and Schaffer

Use square-root LASSO:

```
rlasso y p*,sqrt partial(x)
rlasso s p*,sqrt partial(x)
```

Use loop over macro `e(selected)` to select terms

Improvement 2: Yes it Works!

Table 1: Monte Carlo Output: 5 Sectors

	$\tau_c = 1$		$\tau_c = \beta_c$		$\tau_1 \neq 1$	
	RMSE	Bias	RMSE	Bias	RMSE	Bias
			N=1000			
OLS	0.060	-0.046	0.112	-0.105	0.064	-0.051
Dahl P1	0.049	-0.027	0.087	-0.077	0.062	-0.048
Full	0.064	0.003	0.067	-0.024	0.069	-0.037
LASSO	0.056	0.010	0.060	-0.018	0.058	-0.029
			N=10000			
OLS	0.048	-0.046	0.105	-0.105	0.052	-0.051
Dahl P1	0.019	-0.013	0.055	-0.054	0.045	-0.044
Full	0.037	0.014	0.034	0.004	0.035	-0.018
LASSO	0.034	0.018	0.032	0.014	0.027	-0.009

Empirical Example

The Returns to Schooling

Sample: white males, 25-54, using 1990 US Census.

Migration probabilities estimated using:

- Birth state
- 5 education categories
- Married
- # children 5-18, # children <5
- Divorced
- Live with roommate, family member, alone

Final Results

Table 2: Corrected Estimates versus OLS

	Calif.	Florida	Illinois	Kansas	NY	Texas
			OLS			
College	0.4291	0.4506	0.3689	0.3465	0.4399	0.5166
	(0.0075)	(0.0098)	(0.0096)	(0.0192)	(0.0084)	(0.0086)
Adv	0.5865	0.6618	0.5445	0.4970	0.6037	0.6840
	(0.0105)	(0.0154)	(0.0138)	(0.0315)	(0.0113)	(0.0131)
			Double-Post LASSO			
College	0.3727	0.3919	0.3779	0.3737	0.4192	0.5036
	(0.0138)	(0.0145)	(0.0233)	(0.0345)	(0.0248)	(0.0167)
Adv	0.4864	0.5344	0.4798	0.4807	0.5462	0.6727
	(0.0205)	(0.0209)	(0.023)	(0.0447)	(0.0145)	(0.019)

Final Results

Table 3: Hausman Test of Difference

	Calif.	Florida	Illinois	Kansas	NY	Texas
			LASSO v OLS			
College	-5.586^{***}	-5.823^{***}	0.955	2.763	-2.032	-1.254
Adv	-10.686^{***}	-13.021^{***}	-7.042^{***}	-2.187	-6.185^{***}	-1.5
			LASSO v Dahl			
College	-5.146^{***}	-4.489^{***}	4.854^{**}	2.809	7.366^{***}	0.727
Adv	-8.294^{***}	-11.12^{***}	-1.507	-1.648	4.893^{***}	2.334