Using the stepwise Neyman-orthogonal estimators in Stata

David M. Drukker — Sam Houston State University
Di Liu — Stata

Canadian Stata Conference
23 September 2021

Overview

- This talk is about methods and software for estimating the causal impact of a few covariates on an outcome in a sparse high-dimensional model
- This talk
  - defines sparse high-dimensional models
  - discusses Neyman-orthogonal (NO) estimators for the parameters of interest
  - discusses why use BIC-based stepwise instead of the lasso for an NO estimator
  - discusses the the `swpo` command

What's a high dimensional model?

- I have an extract of the data Sunyer et al. (2017) used to estimate the effect air pollution on the response time of primary school children

$$\mathbf{E}[\texttt{htime}_i | \texttt{no2\_class}, \mathbf{x}] = \exp(\texttt{no2\_class}_i \gamma + \mathbf{x}_i \boldsymbol{\beta})$$

`htime`	the response time on test of child i (hit time)
`no2_class`	air-pollution level in the school of child i
$\mathbf{x}_i$	vector of control variables that might need to be included

- I want to estimate the effect `no2_class` on `htime` and a confidence interval for the size of this effect

High-dimensional models for inference

$$\mathbf{E}[\mathtt{htime}_i|\mathtt{no2\_class},\mathbf{x}] = \exp(\mathtt{no2\_class}_i\gamma + \mathbf{x}_i\boldsymbol{\beta})$$

- If the number of covariates in $\mathbf{x}$ is small relative to the number of observations
  - I can simply include all the controls in $\mathbf{x}$
- In **high-dimensional** models, there are too many potential control covariates in $\mathbf{x}$ to reliably estimate γ when all the controls are included
- There are 252 controls in $\mathbf{x}$, but I only have 1,036 observations
- I cannot reliably estimate γ if I include all 252 controls

Potential solutions

$$\mathbf{E}[\texttt{htime}_i|\texttt{no2\_class}, \mathbf{x}] = \exp(\texttt{no2\_class}_i\gamma + \mathbf{x}_i\boldsymbol{\beta})$$

- Suppose that $\tilde{\mathbf{x}}$ contains the subset of $\mathbf{x}$ that must be included to get a good estimate of γ for the sample size that I have
- If I knew $\tilde{\mathbf{x}}$, I could use the model

$$\mathbf{E}[\texttt{htime}_i|\texttt{no2\_class}, \mathbf{x}] = \exp(\texttt{no2\_class}_i\gamma + \tilde{\mathbf{x}}_i\boldsymbol{\beta})$$

  - I am willing to assume the number of variables in $\tilde{\mathbf{x}}_i$ is small relative to the sample size
    - This is a **sparsity** assumption

- A **high-dimentional** model is one in which there are too many potential covariates, given the sample size
- A **sparse** high-dimentional model in one in which, we only need to include a few of the many potential covariates
  - Few is defined relative to the sample size
- We must solve two problems to do estimation and inference in a sparse high-dimensional model
  1. How to select the few important covariates?
  2. How to get an estimator that is robust to the first stage covariate selection

Theory-based model selection

- The traditional approach would be to use theory to determine which covariates should be included
  - Theory tells us to include controls $\check{\mathbf{x}}$
- Poisson quasi maximum likelihood (QML) of `htime` on `no2_class` and controls $\check{\mathbf{x}}$
  - Let $\widehat{\gamma}_{\check{\mathbf{x}}}$ be estimator with theory-based controls
  - Let $\widehat{\gamma}_{\tilde{\mathbf{x}}}$ be estimator with best-approximating-model controls
  - $\widehat{\gamma}_{\tilde{\mathbf{x}}}$ converges to γ but $\widehat{\gamma}_{\check{\mathbf{x}}}$ does not converge to γ
    - Live with large-sample bias from theory-based covariate selection

- Many researchers want to use the lasso and other data-based methods to perform the covariate selection
  - These methods should be able to remove the large-sample bias arising from theory-based covariate selection
- Some post-covariate-selection estimators provide reliable inference for the few parameters of interest

  Some do not

A naive approach

- Naive estimator:
  1. Use covariate-selection to obtain estimate of which covariates in $\mathbf{x}$ are in $\tilde{\mathbf{x}}$
     Denote estimate by `xhat`
  2. Use QML Poisson to estimate γ and $\tilde{\beta}$
     `poisson htime no2_class xhat`

Why naive approach fails

- Unfortunately, naive estimators that use the selected covariates as if they were $\tilde{\mathbf{x}}$ provide unreliable inference in repeated samples
  - Covariate-selection methods make too many mistakes in estimating $\mathbf{x}$ when some of the coefficients are small in magnitude
  - Here is an example of small coefficient
    - A nonzero coefficient with a magnitude between 1 and 3 times its standard error is small
  - If your model only approximates the process that generated the data, there are approximation terms
    - The coefficients on some of the approximating terms are probably small
- See Leeb and Pötscher (2005), Leeb and Pötscher (2006), Leeb and Pötscher (2008), and Pötscher and Leeb (2009)

Missing small-coefficients covariates matters

- It might seem that not finding covariates with small coefficients does not matter
  - But it does
- When some of the covariates have small coefficients, the distribution of the covariate-selection method is not sufficiently concentrated on the set of covariates that best approximates the process that generated the data
  - Covariate-selection methods will frequently miss the covariates with small coefficients causing omitted variable bias
- The random inclusion or exclusion of these covariates causes
  - the distribution of the naive post-selection estimator to be not normal
  - it makes the usual large-sample theory approximation invalid in theory and unreliable in finite samples

Let's get specific

- The regression function is

$$\mathbf{E}[y|\mathbf{d},\mathbf{x}] = \exp(\mathbf{d}\boldsymbol{\alpha}' + \tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}}') \tag{1}$$

  where

  - **d** includes the few covariates of interest
  - $\tilde{\mathbf{x}}$ is the subset of **x** that belong in the model
    - there are too many covariates in **x** to use the quasi-maximum-likelihood (QML) Poisson estimator for the model

    $$\mathbf{E}[y|\mathbf{d},\mathbf{x}] = \exp(\mathbf{d}\boldsymbol{\alpha}' + \mathbf{x}\boldsymbol{\beta}')$$

    - If you knew the subset $\tilde{\mathbf{x}}$ you could estimate $\boldsymbol{\alpha}$ and the $\boldsymbol{\beta}$ the model in (1)

$$\mathbf{E}[y|\mathbf{d},\mathbf{x}] = \exp(\mathbf{d}\boldsymbol{\alpha}' + \tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}}')$$

A series of seminal papers

- Belloni, Chen, Chernozhukov, and Hansen (2012);
- Belloni, Chernozhukov, and Hansen (2014); and
- Belloni, Chernozhukov, and Wei (2016)

  derived a series of Neyman-orthogonal estimators that provide reliable inference about $\boldsymbol{\alpha}$
- These estimators use a covariate-selection method to select $\tilde{\mathbf{x}}$
- The cost of using a covariate-selection method is that these Neyman-orthogonal estimators do not produce estimates for $\tilde{\boldsymbol{\beta}}$

- When you use two-step estimators, you usually have to adjust your standard errors to account for the parameters you estimated in the first step
  - When you estimate average partial effects, you have to adjust for estimating the coefficients in the first stage
  - Stack the moment conditions
- When you
  1. do model selection
  2. use the selected model

  you have to use an estimator in the second stage that is robust to the model selection mistakes made in the first step

  An NO estimator uses moment equations that have had the effect of the first stage model selection removed

- In a linear model NO estimators end up being an extension of the partialing out algorighm we all learned in first regression class
  - Stata calls NO estimators partialling-out estimators
- NO algorithm for

$$y_i = d_i\gamma + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$$

1. Use selection method to find $\mathbf{x}_y$ (subset of $\mathbf{x}$) that should be included in model for y
2. Let $\tilde{y}$ be residuals from regressing y on $\mathbf{x}_y$
3. Use selection method to find $\mathbf{x}_d$ (subset of $\mathbf{x}$) that should be included in model for d
4. Let $\tilde{d}$ be residuals from regressing d on $\mathbf{x}_d$
5. Estimate γ from OLS of $\tilde{y}_x$ on $\tilde{y}_d$

Covariate selection

- Methods for covariate selection
  - Best subset regression
    - Compute the BIC, or another IC, for all possible subsets of $\mathbf{x}$
    - Select the model that minimizes the BIC
    - Infeasible at p gets large, cannot compute all 2^p estimators
  - One can view the lasso as a feasible convex optimization problem that approximates the best-subset problem
    - The lasso has tuning parameters that must be selected
    - Each method of selecting the lasso tuning parameters is, in effect, a different version of the lasso
  - Stepwise algorithms are another way to approximate the best-subset problem

- Belloni, Chernozhukov, Hansen and coathors use a particular version of the least absolute shrinkage and selection operator (lasso) to perform covariate selection
    - See Hastie et al. (2015) and Belloni et al. (2012) for introductions to the lasso and the form used by Belloni, Chernozhukov, Hansen and coathors
- In our papers, we look at using different versions of the lasso and at using BIC-based stepwise

- BIC based stepwise algorithm

  1. Let $\mathbf{x}_f$ be the full set of potential covariates
  2. Let $\mathbf{x}_{in}$ be the covariates to include in the model
     - At the start let $\mathbf{x}_{in}$ include the constant term
  3. Let BIC_c be the BIC for the current model of QML of y on $\mathbf{x}_{in}$
  4. For each covariate j in $\mathbf{x}_f$, let BIC_j be the for the model of y on $\mathbf{x}_{in}$ and x_j
  5. Let $\tilde{j}$ the j that yields the smallest BIC_j
  6. If $\text{BIC}_{\tilde{j}} < \text{BIC}_c$, then
     - add $x_{\tilde{j}}$ to $\mathbf{x}_{in}$
     - remove $x_{\tilde{j}}$ from $\mathbf{x}_f$
     - let $\text{BIC}_c = \text{BIC}_{\tilde{j}}$
     - go to step 4

     else
     exit

- See Drukker and Liu (2021) and citations therein for more details

Why consider forward stepwise

- Drukker and Liu (2021)
  - discuss a family of data generating processes (DGPs) for which the lasso fails to select the covariates $\tilde{\mathbf{x}}$ in finite samples
  - present simulation evidence that a BIC-based forward stepwise method **can** reliably select the $\tilde{\mathbf{x}}$ from $\mathbf{x}$ for DGPs in this family
  - present simulation evidence that a testing-based forward stepwise method **cannot** reliably select the $\tilde{\mathbf{x}}$ from $\mathbf{x}$ for DGPs in this family
- Using a BIC-based forward stepwise method takes longer than lasso-based methods

  Can take **much** longer

  You are trading time for selection accuracy for some DGPs

- Iterated sure independence screening (SIS) uses a first step that removes variables that have no marginal predictive power. The iterative process puts back the variables that have conditional predictive power and removes the ones that were false included in the first step.
- We are currently looking into using a version of iterated SIS to reduce the computation time of BIC-based forward-stepwise NO estimators
    - Fan and Lv (2008), Fan et al. (2009), and Fan and Song (2010) provide introductions to iterative SIS

Use extract of data from Sunyer et al. (2017)

```
. use breathe7, clear
. describe
Contains data from breathe7.dta
 Observations:          1,089
    Variables:             20                  22 Sep 2021 14:39
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
htime           double  %10.0g                ANT: mean hit reaction time (ms)
no2_class       float   %9.0g                 Classroom NO2 levels (g/m3)
sev_sch         float   %9.0g                 School vulnerability index
noise_sch       float   %9.0g                 Measured school noise (in dB)
age             float   %9.0g                 Child´s age (in years)
ppt             double  %10.0g                Daily total precipitation
grade           byte    %9.0g      grade      Grade in school
sex             byte    %9.0g      sex        Sex
age_start_sch   double  %4.1f                 Age started school
oldsibl         byte    %1.0f                 Older siblings living in house
youngsibl       byte    %1.0f                 Younger siblings living in house
lbfeed          byte    %19.0f     bfeed      duration of breastfeeding
smokep          byte    %3.0f      noyes      1 if smoked during pregnancy
feduc4          byte    %17.0g     edu        Paternal education
meduc4          byte    %17.0g     edu        Maternal education
sev_home        float   %9.0g                 Home vulnerability index
no2_home        float   %9.0g                 Residential NO2 levels (g/m3)
overwt_who      byte    %32.0g     over_wt    WHO/CDC-overweight 0:no/1:yes
ndvi_mn         double  %10.0g                Home greenness (NDVI), 300m
                                                buffer
lbweight        float   %9.0g                 1 if low birthweight
--------------------------------------------------------------------------------
Sorted by:
```

Potential Controls I

```
.
. local ccontrols "sev_home sev_sch age no2_home ppt ndvi_mn noise_sch"
.
. local fcontrols "grade sex meduc4 "
.
. local allcontrols "c.(`ccontrols´) i.(`fcontrols´) "
. local allcontrols "`allcontrols´ i.(`fcontrols´)#c.(`ccontrols´) "
```

BIC-stepwise-based results

```
. posw htime no2_class, controls(`allcontrols´) model(poisson) method(bic)
select controls for htime using stepwise bic
select controls for no2_class using stepwise bic
Partialing-out stepwise bic          Number of obs                 =      1,084
                                     Number of controls            =         79
                                     Number of selected controls   =         45
                                     Wald chi2(1)                  =      30.92
Model: poisson                       Prob > chi2                   =     0.0000
```

htime	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
no2_class	.0034337	.0006175	5.56	0.000	.0022234	.0046439

```
Note: Chi-squared test is a Wald test of the coefficients of the variables of
      interest jointly equal to zero.
. nlcom exp(_b[no2_class])
       _nl_1: exp(_b[no2_class])
```

htime	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	1.00344	.0006196	1619.47	0.000	1.002225	1.004654

Another microgram of NO2 per cubic meter increases the mean reaction time by about 0.3%

lasso-based results

```
. popoisson htime no2_class, controls(`allcontrols´) coef
Estimating lasso for htime using plugin
Estimating lasso for no2_class using plugin
Partialing-out Poisson model          Number of obs                =      1,084
                                      Number of controls           =         79
                                      Number of selected controls  =         10
                                      Wald chi2(1)                 =      29.40
                                      Prob > chi2                  =     0.0000
```

htime	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
no2_class	.0032534	.0006	5.42	0.000	.0020773	.0044294

```
Note: Chi-squared test is a Wald test of the coefficients of the variables
      of interest jointly equal to zero. Lassos select controls for model
      estimation. Type lassoinfo to see number of selected variables in each
      lasso.
. nlcom exp(_b[no2_class])
       _nl_1: exp(_b[no2_class])
```

htime	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	1.003259	.000602	1666.56	0.000	1.002079	1.004439

Another microgram of NO2 per cubic meter increases the mean ...ion time by about 0.3%

Conclusions

- So far
  - Sparse high-dimensional models require covariate selection
  - You must use an NO estimator to account for covariate selection
  - There are DGPs for which an NO estimator that uses BIC-stepwise will perform well, but an NO estimator that uses lasso will not perform well
- Future
  - Use iterated SIS combined with BIC stepwise to get dramatically faster but just as accurate results

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6): 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81(2): 608–650.

Belloni, A., V. Chernozhukov, and Y. Wei. 2016. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics* 34(4): 606–619.

Drukker, D., and D. Liu. 2021. Finite-sample results for lasso and stepwise Neyman-orthogonal Poisson estimators. *Under review at Econometric Reviews* .

Fan, J., and J. Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5): 849–911.

Fan, J., R. Samworth, and Y. Wu. 2009. Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research* 10: 2013–2038.

Fan, J., and R. Song. 2010. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* 38(6): 3567–3604.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Rotaon: CRC Press.

Leeb, H., and B. Pötscher. 2005. Model Selection and Inference: Facts and Fiction. *Econometric Theory* 21: 21–59.

Leeb, H., and B. M. Pötscher. 2006. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics* 34(5): 2554–2591.

———. 2008. Sparse estimators and the oracle property, or the return of Hodges estimator. *Journal of Econometrics* 142(1): 201–211.

Pötscher, B. M., and H. Leeb. 2009. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis* 100(9): 2065–2082.

Sunyer, J., E. Suades-Gonzlez, R. Garca-Esteban, I. Rivas, J. Pujol, M. Alvarez-Pedrerol, J. Forns, X. Querol, and X. Basagaa. 2017. Traffic-related Air Pollution and Attention in Primary School Children: Short-term Association. *Epidemiology* 28(2): 181–189.