Inference for parameters of interest after lasso model selection

David M. Drukker

Executive Director of Econometrics
Stata

Canadian Stata Users Group meeting
25 May 2019

- High-dimensional models include too many potential covariates for a given sample size
- 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

$$htime_i = no2_i\gamma + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$$

$htime$	measure of the response time on test of child i (hit time)
$no2$	measure of the polution level in the school of child i
$\mathbf{x}_i$	vector of control variables that might need to be included

- There are 252 controls in $\mathbf{x}$, but I only have 1,084 observations
- I cannot reliably estimate γ if I include all 252 controls

Potential solutions

$$htime_i = no2_i\gamma + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$$

- I am willing to believe that the number of controls that I need to include is small relative to the sample size
    - This is known as a sparsity assumption
- 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

$$htime_i = no2_i\gamma + \tilde{\mathbf{x}}_i\tilde{\boldsymbol{\beta}} + \epsilon_i$$

So, the problem is that I don't know which variables belong in $\tilde{\mathbf{x}}$ and which do not

$$htime_i = no2_i\gamma + \tilde{\mathbf{x}}_i\tilde{\boldsymbol{\beta}} + \epsilon_i$$

- Now I have a covariate-selection problem
  - Which of the controls in $\mathbf{x}$ belong in $\tilde{\mathbf{x}}$?
- Historically, I would use theory to decide which variables go into $\tilde{\mathbf{x}}$
- Many researchers want to use data-based methods or machine-learning methods to perform the covariate selection
- Some post-covariate-selection estimators provide reliable inference for the few parameters of interest

  Some do not

A naive approach

- The "naive" solution is :
  1. Always include the covariates of interest
  2. Use covariate-selection to obtain estimate of which covariates are in $\tilde{\mathbf{x}}$
     Denote estimate by $\widehat{\mathbf{x}}$
  3. Use estimate $\widehat{\mathbf{x}}$ as if it contained the covariates in $\tilde{\mathbf{x}}$
     `regress htime no2 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 coefficient with a magnitude between 1 and 2 times the standard error is small
  - If your model only approximates the functional form of the true model, there are approximation terms
    - The coefficients on some of the approximating terms are most likely small

Missing small-cofficient 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 ommitted variable bias
- The random inclusion or exclusion of these covariates causes the distribution of the naive post-selection estimator to be not normal and makes the usual large-sample theory approximation invalid in theory and unreliable in finite samples

Beta-min condition

- The beta-min condition was invented to rule-out the existence of small coefficients in the model that best approximates the process that generated the data
- Beta-min conditions are super restrictive and are widely viewed as not defensible
    - See Leeb and Potscher (2005), Leeb and Pötscher (2006), Leeb and Pötscher (2008), and Pötscher and Leeb (2009)

Partialing-out estimators

$$htime_i = no2_i\gamma + \tilde{\mathbf{x}}_i\tilde{\beta} + \epsilon_i$$

- A series of seminal papers

  Belloni, Chen, Chernozhukov, and Hansen (2012);
  Belloni, Chernozhukov, and Hansen (2014);
  Belloni, Chernozhukov, and Wei (2016a); and
  Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018)

  derived a series of partialing-out estimators that provide reliable inference for γ

  - These methods use covariate-selection methods to control for $\tilde{\mathbf{x}}$
  - The cost of using covariate-selection methods is that these partialing-out estimators do not produce estimates for $\tilde{\beta}$

Recommendations

- I am going to provide lots of details, but here are two take aways
  1. If you have time, use the cross-fit partialing-out estimator
     - `xporegress`, `xpologit`, `xpopoisson`, `xpoivregress`
  2. If the cross-fit estimator takes too long, use either the partialing-out estimator
     - `poregress`, `pologit`, `popoisson`, `poivregress`

     or the double-selection estimator
     - `dsregress`, `dslogit`, `dspoisson`

```
. use breathe7
.
. local ccontrols "sev_home sev_sch age ppt age_start_sch  oldsibl "
. local ccontrols "`ccontrols´ youngsibl no2_home ndvi_mn noise_sch"
.
. local fcontrols "grade sex lbweight lbfeed smokep "
. local fcontrols "`fcontrols´ feduc4 meduc4 overwt_who"
.
```

```
. describe htime no2_class `fcontrols´ `ccontrols´

                 storage   display    value
variable 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)
grade           byte     %9.0g      grade      Grade in school
sex             byte     %9.0g      sex        Sex
lbweight        float    %9.0g                 1 if low birthweight
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
overwt_who      byte     %32.0g     over_wt    WHO/CDC-overweight 0:no/1:yes
sev_home        float    %9.0g                 Home vulnerability index
sev_sch         float    %9.0g                 School vulnerability index
age             float    %9.0g                 Child´s age (in years)
ppt             double   %10.0g                Daily total precipitation
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
no2_home        float    %9.0g                 Residential NO2 levels (g/m3)
ndvi_mn         double   %10.0g                Home greenness (NDVI), 300m
                                                 buffer
noise_sch       float    %9.0g                 Measured school noise (in dB)
```

```
. xporegress htime no2_class, controls(i.(`fcontrols´) c.(`ccontrols´)  ///
>                 i.(`fcontrols´)#c.(`ccontrols´))
Cross-fit fold 1 of 10 ...
Estimating lasso for htime using plugin
Estimating lasso for no2_class using plugin
```

(*output omitted*)

```
Cross-fit fold 10 of 10 ...
Estimating lasso for htime using plugin
Estimating lasso for no2_class using plugin
Cross-fit partialed-out                Number of obs                =      1,084
linear model                           Number of controls           =        252
                                       Number of selected controls  =         15
                                       Number of folds in cross-fit =         10
                                       Number of resamples          =          1
                                       Wald chi2(1)                 =      25.36
                                       Prob > chi2                  =     0.0000
```

htime	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
no2_class	2.353006	.4672161	5.04	0.000	1.437279	3.268732

```
Note: Chi-squared test is a Wald test of the coefficients of the variables of
interest jointly equal to zero.
```

Another microgram of NO2 per cubic meter increases the mean reaction time by 2.35 milliseconds.

```
. poregress htime no2_class, controls(i.(`fcontrols´) c.(`ccontrols´)  ///
>                i.(`fcontrols´)#c.(`ccontrols´))
Estimating lasso for htime using plugin
Estimating lasso for no2_class using plugin
Partialed-out linear model          Number of obs               =      1,084
                                    Number of controls          =        252
                                    Number of selected controls =         11
                                    Wald chi2(1)                =      24.45
                                    Prob > chi2                 =     0.0000
```

htime	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
no2_class	2.286149	.4623136	4.95	0.000	1.380031	3.192267

```
Note: Chi-squared test is a Wald test of the coefficients of the variables of
interest jointly equal to zero.
```

Another microgram of NO2 per cubic meter increases the mean reaction time by 2.29 milliseconds.

Estimators

- Describe estimators implemented in `poregress`, and `xporegress`
- Estimators use the least absolute shrinkage and selection operator (lasso) to perform covariate-selection
    - I discuss lasso details after describing estimators
    - For now just think of lasso as covariate-selection method that works when the number of potential covariates is large

      The number of potential covariates p can be greater than the number of observations N

Partialing-out estimator for linear model

- Consider model

$$y = d\gamma + \mathbf{x}\boldsymbol{\beta} + \epsilon$$

- For simplicity, d is a single variable, all methods handle multiple variables
- I discuss a linear model
    - Nonlinear models have similar methods that involve more details

PO estimator for linear model (I)

$$y = d\gamma + \mathbf{x}\boldsymbol{\beta} + \epsilon$$

1. Use a lasso of y on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_y$ that predict y
2. Regress y on $\tilde{\mathbf{x}}_y$ and let $\tilde{y}$ be residuals from this regression
3. Use a lasso of d on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_d$ that predict d
4. Regress d on $\tilde{\mathbf{x}}_d$ and let $\tilde{d}$ be residuals from this regression
5. Regress $\tilde{y}$ on $\tilde{d}$ to get estimate and standard error for γ

- Only the coefficient on d is estimated
- Not estimating $\boldsymbol{\beta}$ can be viewed as the cost of getting reliable estimates of γ that are robust to the mistakes that model-selection techniques make

PO estimator for linear model (II)

$$y = d\gamma + \mathbf{x}\boldsymbol{\beta} + \epsilon$$

1. Use a lasso of y on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_y$ that predict y
2. Regress y on $\tilde{\mathbf{x}}_y$ and let $\tilde{y}$ be residuals from this regression
3. Use a lasso of d on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_d$ that predict d
4. Regress d on $\tilde{\mathbf{x}}_d$ and let $\tilde{d}$ be residuals from this regression
5. Regress $\tilde{y}$ on $\tilde{d}$ to get estimate and standard error for γ

- This is an extension of the partialing-out method for obtaining the ordinary least squares (OLS) estimate for the coefficient and standard error on d (Also known as the result of the Frisch-Waugh-Lovell theorem)

$$y = d\gamma + \mathbf{x}\boldsymbol{\beta} + \epsilon$$

1. Use a lasso of y on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_y$ that predict y
2. Regress y on $\tilde{\mathbf{x}}_y$ and let $\tilde{y}$ be residuals from this regression
3. Use a lasso of d on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_d$ that predict d
4. Regress d on $\tilde{\mathbf{x}}_d$ and let $\tilde{d}$ be residuals from this regression
5. Regress $\tilde{y}$ on $\tilde{d}$ to get estimate and standard error for γ

- Heuristically, the moment conditions used in step 5 are unrelated to the selected covariates
- Formally, the moments conditions used in step 5 have been orthogonalized, or "immunized" to small mistakes in covariate selection
    - Chernozhukov, Hansen, and Spindler (2015a); and Chernozhukov, Hansen, and Spindler (2015b)

Cross-fitting / double-machine-learning PO

- Cross-fitting is also known as double maching learning (DML)
- It uses split-sample techniques on PO estimators
  - to weaken the sparsity condition
  - to get better finite sample performance
- Split-sample techniques further reduce the impact of covariate selection on the estimator for γ

Cross-fitting / double-machine-learning PO

- Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) discusses
  - Why sample-splitting techniques applied to naive machine-learning/covariate-selection estimators do not provide reliable inference inference for γ in repeated samples

    Heuristically, the machine-learning estimators do not converge fast enough to remove the correlation between covariate of interest and the out-of-sample errors in the term predicted by the machine-learning method

Cross-fitting / double-machine-learning PO

- Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) discusses
  - PO estimators simplify the problem and their distributions depend on the correlation between partialed-out covariate of interest and the errors in the term predicted by the machine-learning method
    - Naive estimator depends correlation between the covariate of interest and the errors in the term predicted by the machine-learning method
  - Sample-splitting gets better properties by depending on the out-of-sample correlation between partialed-out covariate of interest and the errors in the term predicted by the machine-learning method instead of the in-sample correlation

1. Split data into samples A and B
2. Using the data in sample A
    1. Use a lasso of y on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_y$ that predict y
    2. Regress y on $\tilde{\mathbf{x}}_y$ and let $\tilde{\boldsymbol{\beta}}_A$ be the estimated coefficients
    3. Use a lasso of d on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_d$ that predict d
    4. Regress d on $\tilde{\mathbf{x}}_d$ and let $\tilde{\boldsymbol{\delta}}_A$ be the estimated coefficients
3. Using the data in sample B
    1. Fill in the residuals for $\tilde{y} = y - \tilde{\mathbf{x}}_y\tilde{\boldsymbol{\beta}}_A$
    2. Fill in the residuals for $\tilde{d} = d - \tilde{\mathbf{x}}_d\tilde{\boldsymbol{\delta}}_A$
4. Using the data in sample B
    1. Use a lasso of y on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_y$ that predict y
    2. Regress y on $\tilde{\mathbf{x}}_y$ and let $\tilde{\boldsymbol{\beta}}_B$ be the estimated coefficients
    3. Use a lasso of d on $\mathbf{x}$ to select covariates $\tilde{\mathbf{x}}_d$ that predict d
    4. Regress d on $\tilde{\mathbf{x}}_d$ and let $\tilde{\boldsymbol{\delta}}_B$ be the estimated coefficients
5. Using the data in sample A
    1. Fill in the residuals for $\tilde{y} = y - \tilde{\mathbf{x}}_y\tilde{\boldsymbol{\beta}}_B$
    2. Fill in the residuals for $\tilde{d} = d - \tilde{\mathbf{x}}_d\tilde{\boldsymbol{\delta}}_B$
6. Regress $\tilde{y}$ on $\tilde{d}$ to get estimates for γ

What's a lasso?

- The linear lasso solves

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ 1/n \sum_{i=1}^{n} (y_i - \mathbf{x}_i \boldsymbol{\beta}') + \lambda \sum_{j=1}^{k} \omega_j |\boldsymbol{\beta}_j| \right\}$$

where

- $\lambda > 0$ is the lasso penalty parameter
- $\mathbf{x}$ contains the p potential covariates
- the ω_j are parameter-level weights known as penalty loadings

What's a lasso?

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ 1/n \sum_{i=1}^{n} (y_i - \mathbf{x}_i \boldsymbol{\beta}') + \lambda \sum_{j=1}^{k} \omega_j |\boldsymbol{\beta}_j| \right\}$$

- As λ grows, the coefficients get "shrunk" towards zero
- The kink in the absolute value function causes some of the elements of $\widehat{\boldsymbol{\beta}}$ to be zero at the solution for some values of λ
- There is a finite value of $\lambda = \lambda_{max}$ for which all the estimated coefficients are zero
- As λ decreases from λ_{max}, the number of nonzero coefficients increases
  - If $p < n$, you obtain the (unpenalized) OLS estimates at $\lambda = 0$

What's a lasso?

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ 1/n \sum_{i=1}^{n} (y_i - \mathbf{x}_i \boldsymbol{\beta}') + \lambda \sum_{j=1}^{k} \omega_j |\boldsymbol{\beta}_j| \right\}$$

- For $\lambda \in (0, \lambda_{max})$ some of the estimated coefficients are exactly zero and some of them are not zero.
  - This is how the lasso works as a covariate-selection method
    - Covariates with estimated coefficients of zero are excluded
    - Covariates with estimated coefficients that not zero are included

Choosing λ

- You must choose λ before you use the lasso to perform covariate selection
- Three methods for selecting λ are
  1. Plug-in estimators
     - These estimators are the default in the PO, DS, and XPO commands
  2. Cross-validation
  3. The adaptive lasso

Plug-in based lasso

- Plug-in estimators find the value of the λ that is large enough to dominate the estimation noise
  - see Belloni, Chernozhukov, and Wei (2016b); Belloni, Chen, Chernozhukov, and Hansen (2012); and Bickel et al. (2009)
  - Belloni, Chernozhukov, and Wei (2016b) and Belloni, Chen, Chernozhukov, and Hansen (2012) show that a lasso with their plug-in estimator achieves an optimal bound on the number of covariates it will include
  - In practice, their bound means that a plug-in-based lasso will include the important covariates and that it will not include many covariates that do not belong in the model

Cross-validated lasso

- Cross-valdiation (CV) finds the $\widehat{\beta}$ that minimizes the out-of-sample prediction error
- CV is widely used, but it is not the best method when using lasso as a covariate-selection method in a PO, XPO, or DS estimator
  - CV tends to choose a λ that causes lasso to include variables whose coefficients are zero in the model that best approximates the true data generating process
  - This over-selection tendency can cause a CV-based PO,DS, XPO estimator to have poor coverage properties

    (Although the XPO estimators are more robust to this problem that PO and DS estimators)

Cross-validated lasso

- See Hastie, Tibshirani, and Wainwright (2015) for lots about how CV lasso is implemented
- See Chetverikov, Liao, and Chernozhukov (2017) for some technical results that could explain the tendency of the cross-validated lasso to include many covariates that do not belong in the model
- See Bühlmann and Van de Geer (2011) for some discussions of the tendency of cross-validated lasso to over select

Adaptive lasso

- The adaptive lasso tends to include more zero-coefficient covariates than a plug-in based lasso and fewer than a cross-validated lasso
- The adaptive lasso is a multistep version of CV
  - The first step is CV
    - The second step does CV among the covariates selected in the first step
    - In the second step, the penalty loadings are set to the inverse of the first-step estimates coefficients
  - Covariate with larger coefficients are more likely to be included in the second step
  - See Zou (2006) and Bühlmann and Van de Geer (2011)

Conclusion

- If you have a model like

$$\mathbf{E}[y|\mathbf{d},\mathbf{x}] = G(\mathbf{d}\gamma + \mathbf{x}\beta)]$$

  where
  - $G()$ is the functional form implied by a linear regression, a logit regression, a Poisson regression
  - $\mathbf{d}$ contains a few know covariates
  - $\mathbf{x}$ contains many potential controls
- You can use
  `xporegress`, `xpologit`, `xpopoisson`, `poregress`, `pologit`, or `popoisson`,
  to estimate γ
- `xpoivregress` and `poivregress` estimate γ for linear models with endogenous covariates when there are many potential instruments and many potential controls

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. 2016a. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics* 34(4): 606–619.

———. 2016b. Post-Selection Inference for Generalized Linear Models With Many Controls. *Journal of Business & Economic Statistics* 34(4): 606–619.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov. 2009. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* 37(4): 1705–1732.

Bühlmann, P., and S. Van de Geer. 2011. *Statistics for*

High-Dimensional Data: Methods, Theory and Applications. Springer Publishing Company, Incorporated.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1): C1–C68.

Chernozhukov, V., C. Hansen, and M. Spindler. 2015a. Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments. *American Economic Review* 105(5): 486–90. URL `http://www.aeaweb.org/articles?id=10.1257/aer.p20151022`.

———. 2015b. Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach. *Annual Review of Economics* 7(1): 649–688.

Chetverikov, D., Z. Liao, and V. Chernozhukov. 2017. On Cross-Validated Lasso. *https://arxiv.org/abs/1605.02214v3* 1–38.

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

Leeb, H., and B. Potscher. 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.

Zou, H. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101(476): 1418–1429.