Algunos estimadores ortogonales de Neyman en Stata

David M. Drukker
Sam Houston State University

Di Liu
Stata

Conferencia Stata México 2021
13 octubre 2021

Plan

- Esta charla trata sobre métodos y software para estimar el impacto causal de algunas covariables en un modelo disperso de alta dimensión
- Esta charla
  - define modelos dispersos de alta dimensión
  - describe los estimadores ortogonales de Neyman (NO) para estimar los parámetros de interés
  - describe por qué utilizar el BIC paso-a-paso en lugar de el lazo para un estimador de NO
    - (Traduczo "forward stepwise" como paso-a-paso)
  - presenta el comando `swpo`

¿Qué es un modelo de alta dimensión?

- Tengo unos de los datos que Sunyer et al. (2017) usaron para estimar el efecto de la contaminación del aire en el tiempo de respuesta de los niños de la escuela primaria

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

`htime`	el tiempo de respuesta en la prueba del niño
`no2_class`	nivel de contaminación del aire en la escuela del niño i
$\mathbf{x}_i$	vector de variables de control que pueden necesitar ser incluido

- Quiero estimar el efecto del `no2_class` en `htime` y quiero un intervalo de confianza válido para el tamaño de este efecto

Modelos de inferencia de alta dimensión

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

- Si el número de covariables en **x** es pequeño relativo al número de observaciones
  - Simplemente puedo incluir todos las variables (controles) en **x**
- En modelos **de alta dimensión**,si se incluye todos las covariables en **x**, el estimador para γ no es confiable
- Hay 252 controles en **x**, pero solo tengo 1036 observaciones
- No puedo estimar de manera confiable γ si incluyo los 252 controles

Soluciones potenciales

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

- Suponga que $\tilde{\mathbf{x}}$ contiene el subconjunto de $\mathbf{x}$ que debe ser incluido para obtener una buena estimación de γ para el tamaño de muestra que tengo
- Si supiera $\tilde{\mathbf{x}}$, podría usar el modelo

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

  - Estoy dispuesto a asumir que el número de variables en $\tilde{\mathbf{x}}_i$ es pequeño en relación con el tamaño de la muestra
    - Esta es una suposición de **sparsity**
      Parece que se dice que el modelo es disperso en español (castellano)

- Un modelo de **alta dimención** es uno en el que hay demasiadas covariables posibles, dado el tamaño de la muestra
- Un modelo **disperso** de alta dimensión es uno en el que, solo hay que incluir **algunos** de las muchas covariables potenciales
  - **algunos** se define en relación con el tamaño de la muestra
- Debemos resolver dos problemas para estimar y sacar inferencia de un modelo dispero de alta dimensión
  1. ¿Cómo seleccionar las pocas covariables importantes?
  2. ¿Cómo obtener un estimador que sea robusto para la primera etapa selección de covariables ?

Selección de modelo basada en la teoría

- El enfoque tradicional sería utilizar la teoría para determinar caules son las que covariables se deben incluir
  - Theory nos dice que incluyamos controles $\check{\mathbf{x}}$
- $\widehat{\gamma}_{\check{\mathbf{x}}}$ es el estimador con controles basados en la teoría
- $\widehat{\gamma}_{\tilde{\mathbf{x}}}$ es el estimador con controles del mejor modelo aproximado
- $\widehat{\gamma}_{\tilde{\mathbf{x}}}$ converge en γ pero $\widehat{\gamma}_{\check{\mathbf{x}}}$ no converge a γ
- Se vive con un sesgo de muestra grande de la selección de covariables basada en la teoría

- Muchos investigadores quieren utilizar el lazo y otros métodos basados en datos para realizar la selección de covariables
    - Estos métodos deberían poder eliminar el sesgo de muestras grandes que surgen de la selección de covariables basada en la teoría
- Algunos estimadores posteriores a la selección de covariables proporcionan inferencias válidas por los parámetros de interés

  Otros no proportionan inferencias válidas

Un enfoque ingenuo

- Estimador ingenuo de
  1. Utilice la selección de covariables para obtener una estimación de cuáles son las covariables de $\mathbf{x}$ que se debe incluir en $\tilde{\mathbf{x}}$
     Denote la estimación por`xhat`
  2. Use QML Poisson para estimar γ y $\tilde{\boldsymbol{\beta}}$
     `poisson htime no2_class xhat`

Por qué falla el enfoque ingenuo

- Los estimadores ingenuos que usan las covariables seleccionadas como si fueran $\tilde{\mathbf{x}}$ proporcionan inferencia inválidas en muestras repetidas
  - Los métodos de selección de covariables cometen demasiados errores en estimar $\mathbf{x}$ cuando algunos de los coeficientes son pequeños en magnitud
  - Aquí hay un ejemplo de un coeficiente pequeño
    - Un coeficiente distinto de cero con una magnitud entre 1 y 3 veces su error estándar es pequeño
  - Si su modelo solo se aproxima al proceso que generó los datos, hay términos de aproximación
    - Los coeficientes de algunos de los términos aproximados son probablemente pequeño
- Vea Leeb and Pötscher (2005), Leeb and Pötscher (2006), Leeb and Pötscher (2008) y Pötscher and Leeb (2009)

- Puede parecer que no importa no encontrar covariables con coeficientes pequeños
  - Pero sí importa
- Cuando algunas de las covariables tienen coeficientes pequeños, la distribución del método de selección de covariables no se concentra suficientemente en el conjunto de covariables que mejor se aproxima al proceso que generó los datos
  - Los métodos de selección de covariables frecuentemente faltan covariables con coeficientes pequeños
  - La falta de estas covariables causa un poco de sesgo de variable omitida
- La inclusión o exclusión aleatoria de estas covariables
  - provocala distribución del estimador de post-selección ingenuo para ser no normal
- Usar la invalida aproximación normal proporciona inferencias inválidas en muestras finitas

Seamos específicos

- La función de regresión es

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

donde

  - $\mathbf{d}$ incluye las pocas covariables de interés
  - $\tilde{\mathbf{x}}$ es el subconjunto de $\mathbf{x}$ que pertenecen a el modelo
    - hay demasiadas covariables en $\mathbf{x}$ para usar el estimador de Poisson de cuasi-máxima verosimilitud (QML) para modelo

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

    - Si conociera el subconjunto $\tilde{\mathbf{x}}$ podría estimar $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ el modelo en (1)

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

Una serie de artículos seminales

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

  derivó una serie de estimadores ortogonales de Neyman que proporcionan inferencia sobre $\boldsymbol{\alpha}$
- Estos estimadores utilizan un método de selección de covariables para seleccionar $\tilde{\mathbf{x}}$
- El costo de usar un método de selección de covariables es que estos estimadores ortogonales de Neyman no producen estimaciones para $\tilde{\boldsymbol{\beta}}$

- Cuando se utiliza estimadores de dos pasos, normalmente uno tiene que ajustar sus errores estándar para tomar en cuenta las estimaciones del primer paso
  - Cuando estima el promedio de los efectos parciales, un tiene que ajustar los errores estandares por la estimaciones de los coeficientes en la primera etapa
  - Apila las condiciones del momento
- Cuando tu
  1. escojes las covariables
  2. y usas la covariable seleccionada en un modelo

  tienes que usar un estimador en la segunda etapa que sea robusto a los errores de selección cometidos en el primer paso

  Un estimador NO usa ecuaciones de momento que son robustas a las selecciones en la primera etapa

- En un modelo lineal, los estimadores de NO terminan siendo una extensión del parcialización que todos aprendimos en la primera clase de regresión
  - Stata llama los estimadores NO "partialling out"
- El algoritmo NO para

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

1. Use el método de selección para encontrar $\mathbf{x}_y$ (subconjunto de $\mathbf{x}$) que debe incluirse en el modelo por y
2. Sea $\tilde{y}$ residuales de la regresión de y en $\mathbf{x}_y$
3. Use el método de selección para encontrar $\mathbf{x}_d$ (subconjunto de $\mathbf{x}$) que debe incluirse en el modelo por d
4. Sea $\tilde{d}$ residuales de la regresión d en $\mathbf{x}_d$
5. Estimación de γ de MCO de $\tilde{y}_x$ en $\tilde{y}_d$

Selección de covariables

- Métodos para la selección de covariables
  - Mejor subconjunto
    - Calcule el BIC, u otro IC, para todos los posibles subconjuntos de $\mathbf{x}$
    - Seleccione el modelo que minimiza el BIC
    - No es factible cuando p se vuelve grande, no se pueden calcular todos los estimadores de 2^p
  - Se puede ver el lazo como un problema de optimización convexa factible que se aproxima el problema del mejor subconjunto
    - El lazo tiene parámetros de ajuste que uno tener que escojer
    - Cada método de escojer los parámetros de ajuste de lazo es, en efecto, una versión del lazo
  - Los algoritmos paso-a-paso son otra forma de aproximar el problema del mejor subconjunto

- Belloni, Chernozhukov, Hansen y coathors utilizar una versión particular del operador de selección y contracción mínima absoluta (lazo) para realizar la selección de covariables
    - Consulte Hastie et al. (2015) y Belloni et al. (2012) para conocer las introducciones lazo y la forma utilizada por Belloni, Chernozhukov, Hansen y coathors
- En nuestros artículos, analizamos el uso de diferentes versiones del lazo y en el uso de BIC paso a paso

¿Qué es un lazo?

- El lazo de Poisson resuelve

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

dónde

  - $\lambda > 0$ es el parámetro de penalización de lazo
  - $\mathbf{x}$ contiene las covariables potenciales p
  - los ω_j son pesos a nivel de parámetro conocidos como penalización cargas
  - λ y ω_j se conocen como parámetros de ajuste de lazo
  - p es el número de covariables potenciales en $\mathbf{x}$ y n es el tamaño de la muestra
  - puede tener p mayor que n

¿Qué es un lazo?

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

- Cuando $\lambda = 0$, se obtiene estimaciones de Poisson QML no penalizadas (cuando $p < n$)
- A medida que crece λ, las estimaciones de las coeficientes se "encogen" hacia cero
  - un cambia sesgo de muestra grande en $\widehat{\boldsymbol{\beta}}$ por un predictor fuera de la muestra que tiene un error cuadrático medio más bajo

¿Qué es un lazo?

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

- La torcedura en la función de valor absoluto hace que algunos de los elementos de $\widehat{\boldsymbol{\beta}}$ sea cero en la solución para algunos valores de λ
  - A medida que crece λ, más estimaciones de coeficientes son exactamente cero
- Hay un valor finito de $\lambda = \lambda_{max}$ para el cual todos los los coeficientes estimados son cero

¿Qué es un lazo?

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

- Para $\lambda \in (0, \lambda_{max})$ algunos de los coeficientes estimados son exactamente cero y algunos de ellos no son cero.
  - Así es como funciona el lazo como método de selección de covariables
    - Las covariables con coeficientes estimados de cero son excluido
    - Covariables con coeficientes estimados distintos de cero están incluidos

Elegir λ

- Debe elegir λ antes de usar el lazo para realizar selección de covariables
- Tres métodos en la literatura para seleccionar λ son
  1. Validación cruzada (CV)
  2. Plugin
  3. Minimizar un criterio de información

- Los resultados teóricos en Chetverikov, Liao, and Chernozhukov (2020) indican que el CV, que es ampliamente utilizado, no debe usarse para estimadores de NO
- Drukker and Liu (2021) presenta evidencia de simulación de que se debe evitar la CV para el lazo para estimadores de NO
- Drukker and Liu (2021) amplía el algoritmo del plugin de modeles lineales a la familia GLM y muestra que funciona bien en simulaciones
    - Este algoritmo se implementa en `lasso` y todos los comandos basados en `lasso` en Stata
- Drukker and Liu (2021) también presenta evidencia de simulación de que el lasso puede producir un estimador de NO que funciona tan bien como el lazo basado en complementos
    - Este algoritmo también se implementa en `lasso` y todos los comandos basados en lazo en Stata

- algoritmo paso a paso basado en BIC
  1. Sea $\mathbf{x}_f$ el conjunto completo de covariables potenciales
  2. Sea $\mathbf{x}_{in}$ las covariables a incluir en el modelo
     - Al principio, $\mathbf{x}_{in}$ incluye el término constante
  3. Sea BIC $_c$ el BIC para el modelo actual de QML de y en $\mathbf{x}_{in}$
  4. Para cada covariable j en $\mathbf{x}_f$, sea BIC $_j$ el para el modelo de y en $\mathbf{x}_{in}$ y x_j
  5. Sea $\tilde{j}$ el j que produce el BIC más pequeño $_j$
  6. Si BIC$_{\tilde{j}} <$ BIC $_c$, entonces
     - agrega $x_{\tilde{j}}$ a $\mathbf{x}_{in}$
     - quitar $x_{\tilde{j}}$ de $\mathbf{x}_f$
     - let BIC $_c =$ BIC $_{\tilde{j}}$
     - vaya al paso 4

  salir
- Consulte Drukker and Liu (2021) y las citas en el mismo para obtener más detalles.

¿Por qué considerar paso a paso?

- Drukker and Liu (2021)
  - discute una familia de procesos de generación de datos (DGP) para lo cual el lazo no puede seleccionar las covariables $\tilde{\mathbf{x}}$ en muestras finitas
  - presenta evidencia de simulación de que un método paso-a-paso basado en BIC **puede** seleccionar $\tilde{\mathbf{x}}$ de $\mathbf{x}$ para los DGP de esta familia
  - presenta evidencia de simulación de que un método paso-a-paso basado en pruebas de hipótesis **no puede** seleccionar $\tilde{\mathbf{x}}$ de $\mathbf{x}$ para los DGP de esta familia
- El uso de un método paso-a-passo basado en BIC lleva más tiempo que los métodos basados en lazo

  Puede tardar **mucho** más

  Es cambiar tiempo por precisión de selección para algunos DGP

- La detección iterada de independencia segura (SIS) utiliza un primer paso que elimina variables que no tienen poder predictivo marginal. El proceso iterativo pone respalda las variables que tienen poder predictivo condicional y elimina las que eran falsos incluidos en el primer paso.
- Actualmente estamos estudiando el uso de una versión de SIS iterado para reducir el tiempo de cálculo de los estimadores de NO escalonados hacia adelante basados en BIC
    - Fan and Lv (2008), Fan et al. (2009) y Fan and Song (2010) proporcionan presentaciones al SIS iterativo

Use unos de datos de Sunyer et al. (2017)

```
 . use breathe7, clear
. describe
Contains data from breathe7.dta
 Observations:         1,089
    Variables:            20                  22 Sep 2021 14:39
```

Variable name	Storage type	Display format	Value 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:
```

Controles potenciales 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´) "
```

Resultados basados en BIC paso a paso

```
. 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

Otro microgramo de NO2 por metro cúbico aumenta la reacción media tiempo en aproximadamente un 0,3 %

resultados basados en lazo

```
. 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

Otro microgramo de NO2 por metro cúbico aumenta la reacción media
no en aproximadamente un 0.3 %

Conclusiones

- Hasta ahora
  - Los modelos dispersos de alta dimensión requieren una selección de covariables
  - Debe utilizar un estimador NO para tener en cuenta la selección de covariables
  - Hay DGP para los que un estimador NO que utiliza BIC paso a paso funcionará bien, pero un estimador de NO que use lazo no funcionará bien
- Futuro
  - Use SIS iterado combinado con BIC paso a paso para ser mucho más rápido pero resultados igualmente precisos

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.

Chetverikov, D., Z. Liao, and V. Chernozhukov. 2020. On Cross-Validated Lasso in High Dimensions. *https://arxiv.org/pdf/1605.02214.pdf* .

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.