Regularización: Lasso y Elastic Net

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Regularization
   - Recap
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - Elastic Net
2. Causal Inference
   - Causality Review: Treatment Effects
   - Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls

Agenda

1. Regularization
   - Recap
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - Elastic Net

2. Causal Inference
   - Causality Review: Treatment Effects
   - Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls

Agenda

Regularización: Motivación

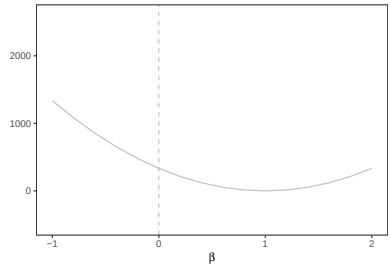
- Las técnicas econometricas estándar no están optimizadas para la predicción porque se enfocan en la insesgadez.
- OLS por ejemplo es el mejor estimador lineal *insesgado*
- OLS minimiza el error "*dentro de muestra*", eligiendo β de forma tal que

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 \tag{1}$$

- pero para predicción, no estamos interesados en hacer un buen trabajo dentro de muestra
- Queremos hacer un buen trabajo, **fuera de muestra**

OLS 1 Dimension

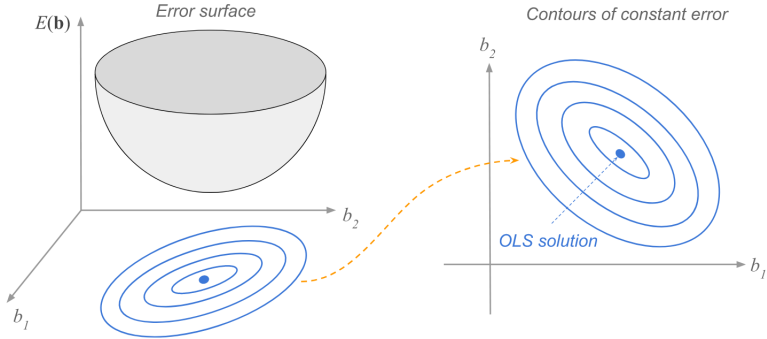
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 \tag{2}$$



App

OLS 2 Dimensiones

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \tag{3}$$



Fuente: https://allmodelsarewrong.github.io

App

Ridge

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Las técnicas de machine learning fueron desarrolladas para hacer este trade-off de forma empírica.
- Vamos a proponer modelos del estilo

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{4}$$

- donde R es un regularizador que penaliza funciones que crean varianza

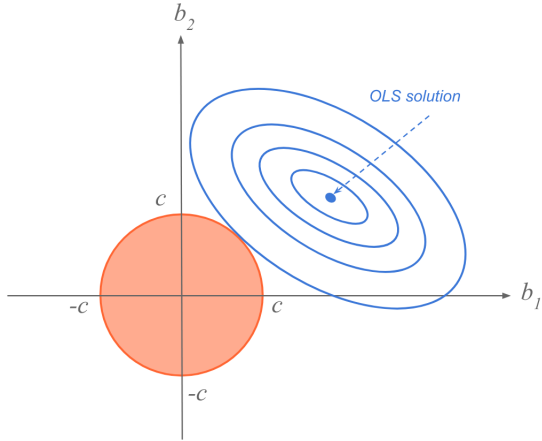
Ridge

- Para un $\lambda \geq 0$ dado, consideremos ahora el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2 \tag{5}$$

Intuición en 2 Dimensiones (Ridge)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2 \right) \leq c \tag{6}$$



Agenda

Lasso

- Para un $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{7}$$

Lasso

- Para un $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{7}$$

  - "LASSO's free lunch": selecciona automáticamente los predictores que van en el modelo ($\beta_j \neq 0$) y los que no ($\beta_j = 0$)
  - Por qué? Los coeficientes que no van son soluciones de esquina
  - $L(\beta)$ es no differentiable

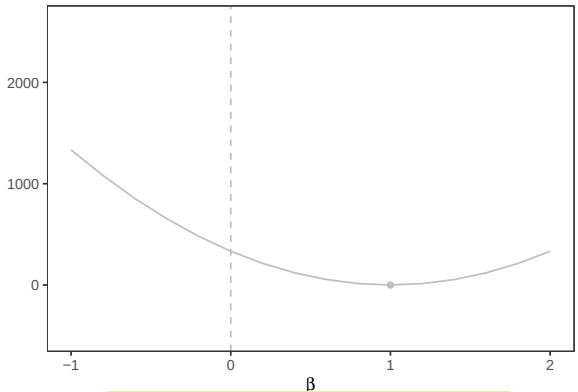
Intuición en 1 Dimensión

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{8}$$

Intuición en 1 Dimensión

$\hat{\beta} > 0$

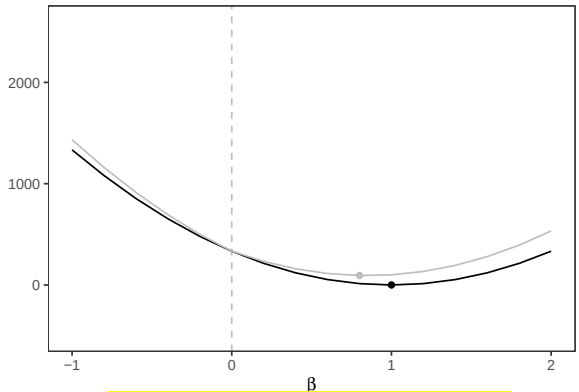
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{9}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

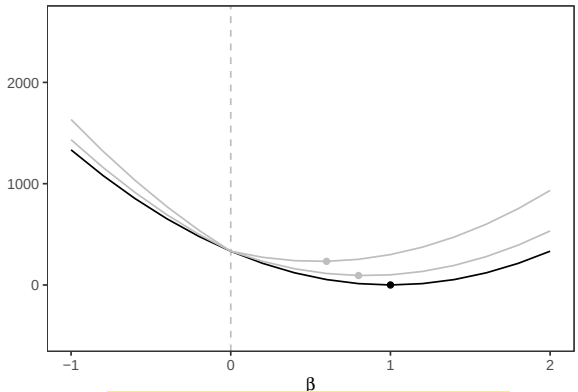
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{10}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

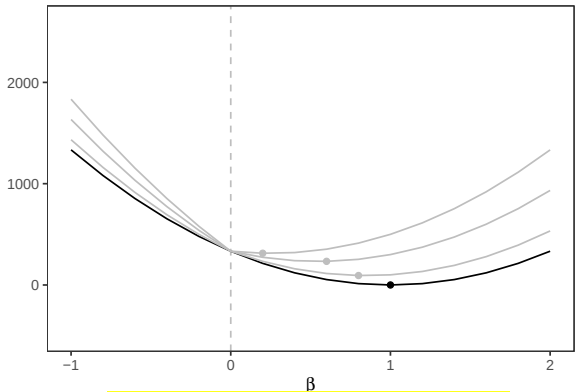
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{11}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

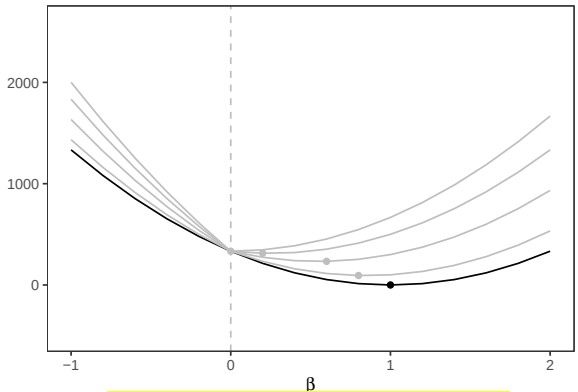
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{12}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

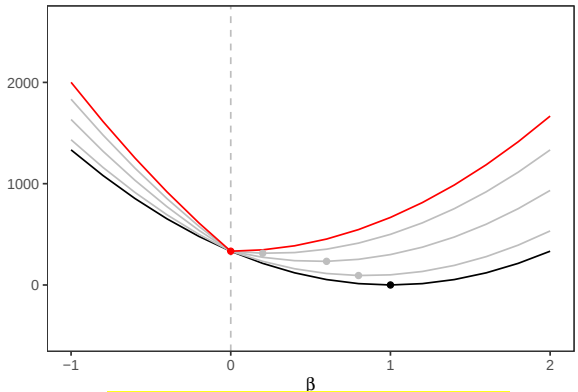
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{13}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

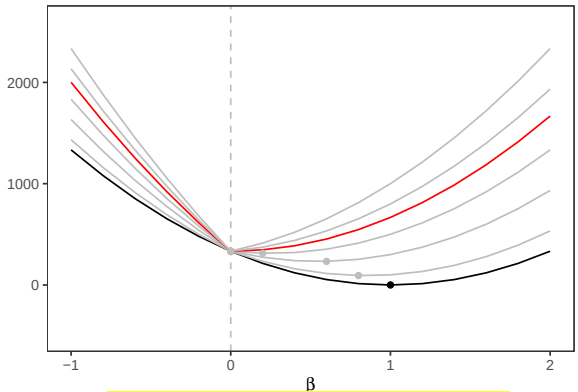
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{14}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{15}$$



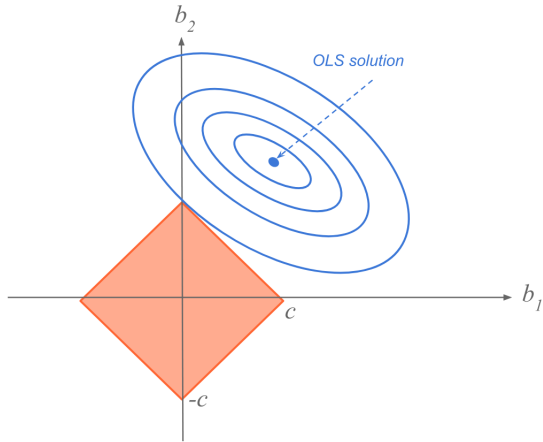
Intuición en 1 Dimension

Solución analitica

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{16}$$

Intuición en 2 Dimensiones (Lasso)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \tag{17}$$



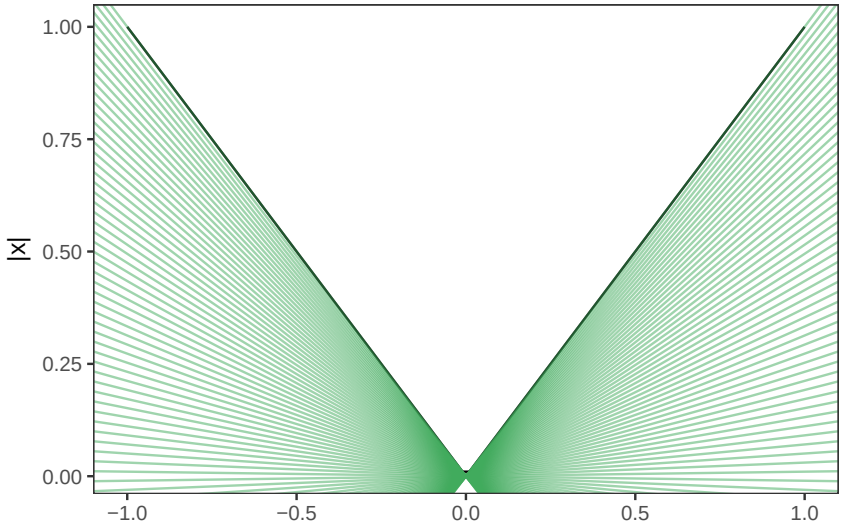
Example

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

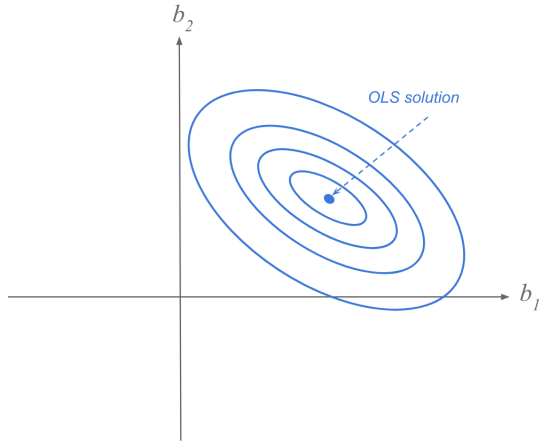
Resumen

- Ridge y Lasso son sesgados, pero las disminuciones en varianza pueden compensar estoy y llevar a un MSE menor
- Lasso encoje a cero, Ridge no tanto
- Importante para aplicación:
  - Estandarizar los datos
  - Elegimos $\lambda \rightarrow$ Validación cruzada

Subgradientes



Coordinate Descent



Agenda

1. Regularization
   - Recap
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - Elastic Net
2. Causal Inference
   - Causality Review: Treatment Effects
   - Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls

Ridge and Lasso: The good and the bad

- Objective 1: Accuracy
  - Minimize prediction error (in one step) $\rightarrow$ Ridge, Lasso
- Objective 2: Dimensionality
  - Reduce the predictor space $\rightarrow$ Lasso's free lunch
- More predictors than observations ($k > n$)
  - OLS fails
  - Ridge augments data
  - Lasso chooses at most n variables

Ridge and Lasso: The good and the bad

OLS when $k > n$

- Rank? Max number of rows or columns that are linearly independent
  - Implies $rank(X_{n\times k}) \leq min(k, n)$
- MCO we need $rank(X_{n\times k}) = k \implies k \leq n$
- If $rank(X_{n\times k}) = k$ then $rank(X'X) = k$
- If $k > n$, then $rank(X'X) \leq n < k$ then $(X'X)$ cannot be inverted
- Ridge works when $k \geq n$

Ridge and Lasso: The good and the bad

Ridge when $k > n$

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} x'_{ij}\beta_j)^2 + \lambda (\sum_{j=1}^{k} \beta_j)^2 \tag{18}$$

- Solution $\rightarrow$ data augmentation
- Intuition: Ridge "adds" k additional points.
- Allows us to "deal" with $k \geq n$

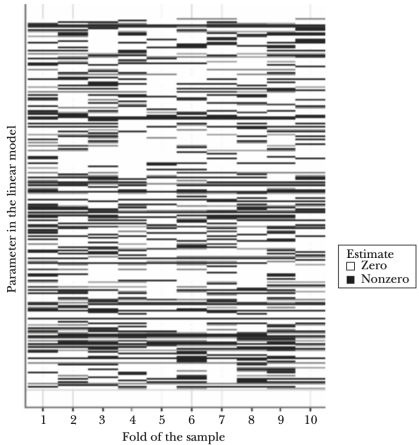
Ridge and Lasso: The good and the bad

Ridge when $k > n$

Ridge and Lasso: The good and the bad

- When we have a group of highly correlated variables,
  - Lasso chooses only one.

Ridge and Lasso: The good and the bad



Ridge and Lasso: The good and the bad

- When we have a group of highly correlated variables,
  - Lasso chooses only one. Makes it unstable for prediction.
  - Ridge shrinks the coefficients of correlated variables toward each other. This makes Ridge "work" better than Lasso. "Work" in terms of prediction error

Agenda

Family of penalized regressions

$$min_{\beta} R(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|^q \tag{19}$$

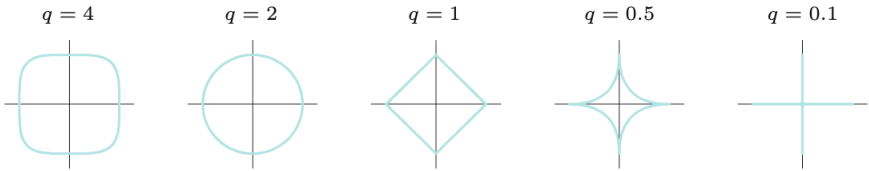


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Agenda

1. Regularization
   - Recap
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - Elastic Net
2. Causal Inference
   - Causality Review: Treatment Effects
   - Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls

Elastic net

$$min_{\beta} EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \tag{20}$$

- Si $\alpha = 1$ Lasso
- Si $\alpha = 0$ Ridge

Elastic Net

- Elastic net: happy medium.
  - Good job at prediction and selecting variables

$$min_{\beta} EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \tag{21}$$

- Mixes Ridge and Lasso
- Lasso selects predictors
- Strict convexity part of the penalty (ridge) solves the grouping instability problem
- How to choose (λ, α)? $\rightarrow$ Bidimensional Crossvalidation
- Recomended lecture: Zou, H. & Hastie, T. (2005)

Example

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

1. Regularization
   - Recap
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - Elastic Net

2. Causal Inference
   - Causality Review: Treatment Effects
   - Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls

Motivation

Motivation

- In this course our objective is prediction
- But since we are economists, inference is always there
- Can we use some of these models to do causal inference?

Agenda

1. Regularization
   - Recap
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - Elastic Net
2. Causal Inference
   - Causality Review: Treatment Effects
   - Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls

Treatment Effects Review

- We observe a sequence of triples $\{(W_i, Y_i, X_i)\}_i^N$,

Treatment Effects Review

Unfortunately, in our data we can only observe one of these two potential outcomes.

Education (X_i)	Treated W_i	No Subsidy $Y_i(0)$	Subsidy $Y_i(1)$	Treatment effect $\tau_i = Y_i(1) - Y_i(0)$
High	1	?	$Y_1(1)$	?
High	0	$Y_2(0)$	?	?
Low	0	$Y_3(0)$	?	?
Low	1	?	$Y_4(1)$	?

Treatment Effects Review

- Before proceeding we need to make a couple of assumptions
- Assumption 1: Unconfoundedness

$$Y_i(1), Y_i(0) \perp W_i \mid X_i \tag{22}$$

- Assumption 2: Overlap

$$\forall\, x \in \text{supp}\,(X), \qquad 0 < P\,(W = 1 \mid X = x) < 1 \tag{23}$$

Average Treatment Effects Review

- Computing the difference for each individual is impossible.
- But we can get the Average Treatment Effect (ATE):

$$\tau := E[Y_i(1) - Y_i(0)] \tag{24}$$

- When our above assumptions are true we have:
- Conditional Average Treatment Effect (CATE)

$$\tau(x) := E[Y_i(1) - Y_i(0)|X_i = x] \tag{25}$$

Agenda

1. Regularization
   - Recap
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - Elastic Net
2. Causal Inference
   - Causality Review: Treatment Effects
   - Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls

Model Selection When the Goal is Causal Inference

Let's start with the following model

$$y_i = \alpha + \beta W_i + g(X_i) + \zeta_i \tag{26}$$

were

- W_i is the treatment/policy variable of interest,
- X_i is a set controls
- $E[\zeta_i | W_i, X_i] = 0$

Model Selection When the Goal is Causal Inference

- Traditional approach: researcher selects X_i
- Problem: mistakes can occur.
- Same if they use an "automatic" model selection approach.
- It can leave out potentially important variables with small coefficients but non zero coefficients out

Model Selection When the Goal is Causal Inference

- The omission of such variables then generally contaminates estimation and inference results based on the selected set of variables. (e.g. OVB)
- The validity of this approach is delicate because it relies on perfect model selection.
- Because model selection mistakes seem inevitable in realistic settings, it is important to develop inference procedures that are robust to such mistakes.
- Solution here: Lasso

Model Selection When the Goal is Causal Inference

- Using Lasso is useful for prediction
- However, naively using Lasso to draw inferences about model parameters can be problematic.
- Part of the difficulty is that these procedures are designed for prediction, not for inference
- Leeb and Pötscher 2008 show that methods that tend to do a good job at prediction can lead to incorrect conclusions when inference is the main objective
- This observation suggests that more desirable inference properties may be obtained if one focuses on model selection over the predictive parts of the economic problem

Agenda

Approximate sparse models

- To fix ideas suppose we have the following model and we want to predict y based on X

$$y_i = g(X_i) + \zeta_i \tag{27}$$

with

  - $E(\zeta_i | g(x_i)) = 0$
  - $i = 1, \ldots, n$ are iid
  - To avoid over-fitting and produce good out of sample prediction we will need to regularize $g(.)$
  - Belloni's et. all approach focuses on an approach that treats $g(X_i)$ as a high-dimensional but that we can approximate linearly

Approximate sparse models

$$g(X_i) = \sum_{j=1}^{p} \beta_j x_{ij} + r_{pi} \tag{28}$$

- where $p >> n$ and r_{pi} is small enough
- Approximate sparsity of this high-dimensional linear model imposes the restriction that linear combinations of only s, where $s << n$; x_{ij} variables provide a good approximation to $g(X_i)$

Approximate sparse models

- We can use Lasso that is slightly modified

$$L(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=2}^{p} |\beta_j|\gamma_j \tag{29}$$

  - where $\lambda > 0$ is the penalty level
  - γ_j are *penalty loadings*
    - *penalty loadings* are chosen to insure equivariance of coefficient estimates to rescaling of x_{ij} and can also be chosen to address heteroskedasticity, clustering, and non-gaussian errors

Agenda

Inference with Selection among Many Controls

- Under the approximate sparse models assumption, we can write our model

$$y_i = \alpha + \beta W_i + X_i'\theta_y + r_{yi} + \zeta_i \tag{30}$$

- where $E[\zeta_i | W_i, x_i, r_{yi}] = 0$
- X_i is a p-dimensional vector with $p >> n$, but approximately sparse
- r_{yi} is an approximation error

Inference with Selection among Many Controls

- How do we proceed?

Inference with Selection among Many Controls

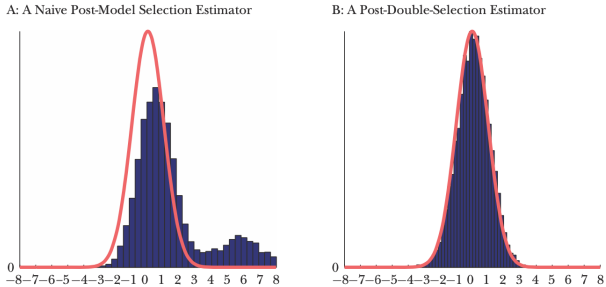
- To prevent model selection mistakes, it is important to consider both equations for selection.
- We apply variable selection methods to each of the two reduced form equations and then use all of the selected controls in estimation of β.
- We select
  1. A set of variables that are useful for predicting y_i, say X_{yi}, and
  2. A set of variables that are useful for predicting W_i, say X_{di}.
- We then estimate β by ordinary least squares regression of y_i on W_i and the union of the variables selected for predicting y_i and W_i, contained in X_{yi} and X_{di}.

Inference with Selection among Many Controls

Figure 1

The "Double Selection" Approach to Estimation and Inference versus a Naive Approach: A Simulation from Belloni, Chernozhukov, and Hansen (forthcoming)

(distributions of estimators from each approach)



Source: Belloni, Chernozhukov, and Hansen (forthcoming).

Notes: The left panel shows the sampling distribution of the estimator of α based on the first naive procedure described in this section: applying LASSO to the equation $y_i = d_i + x_i' \theta_y + r_{yi} + \zeta_i$ while forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. The right panel shows the sampling distribution of the "double selection" estimator (see text for details) as in Belloni, Chernozhukov, and Hansen (forthcoming). The distributions are given for centered and studentized quantities.

Example: Green and Kern, 2012