Lecture 9: Penalizaed Regression

Naijia Liu

Feb. 20 2024

Midterm

- March 7th, Thursday. In class and closed book.
- Gradescope: 11:50am - 1:15pm
- Location: Sever Hall 103 or your own location.
  - We can answer clarification question on Slack and in Sever.
  - **Strongly** recommend you to come in person.
- Concept questions + coding tasks.

Final Project

- Form your own group!

  Find your interest / personality aligned classmates, at most 4, at least 2.

- 1 page write up due on April 5th.

  Research question + Introduce the data.

- First poster draft due on April 12th, Friday

  I will provide you with more examples of how to make a poster.

- Poster final draft deadline on April 18th(non-negotiable, no late submission)

Final Project

- Form your own group!

  Find your interest / personality aligned classmates, at most 4, at least 2.
- 1 page write up due on April 5th.

  Research question + Introduce the data.
- First poster draft due on April 12th, Friday

  I will provide you with more examples of how to make a poster.
- Poster final draft deadline on April 18th(non-negotiable, no late submission)
- Poster Session on April 23rd, Tuesday (usual lecture time).

A World Full of Data

- It is important to prevent political violence.

A World Full of Data

- It is important to prevent political violence.
- Newspaper data everyday everywhere.

A World Full of Data

- It is important to prevent political violence.
- Newspaper data everyday everywhere.
- Can we predict political violence using Newspaper data? (Mueller and Rauh, APSR, 2017)

A World Full of Data

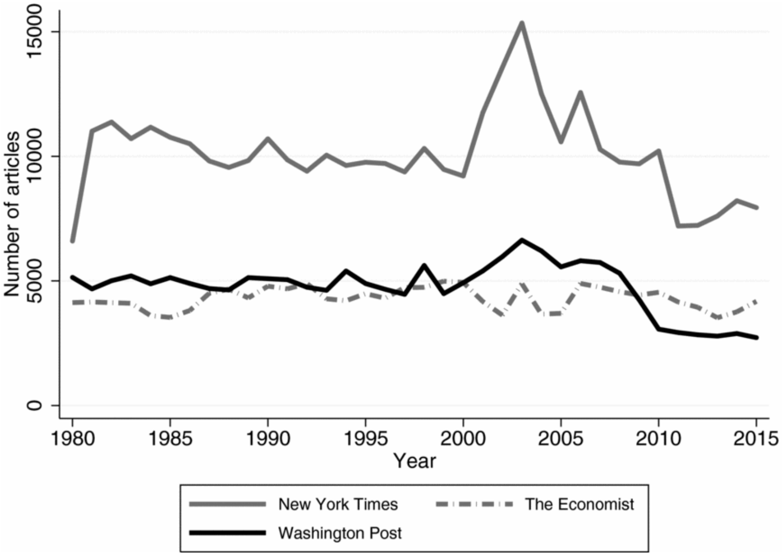
- It is important to prevent political violence.
- Newspaper data everyday everywhere.
- Can we predict political violence using Newspaper data? (Mueller and Rauh, APSR, 2017)
- Within-country variation of newspaper topics is a good predictor of conflict.

A World Full of Data

- It is important to prevent political violence.
- Newspaper data everyday everywhere.
- Can we predict political violence using Newspaper data? (Mueller and Rauh, APSR, 2017)
- Within-country variation of newspaper topics is a good predictor of conflict.
- What are the topics / aspects of newspaper data that we want to use???

$$\text{violence} = \beta_0 + \beta \cdot \text{newspaper features} + \epsilon$$

Newspapers



Number of articles
0
5000
10000
15000
1980
1985
1990
1995
2000
2005
2010
2015
Year
New York Times
The Economist
Washington Post

Data

- All articles on 185 countries from the New York Times (NYT), the Washington Post (WP), and the Economist for all available years since 1975. (700,000 articles in total)

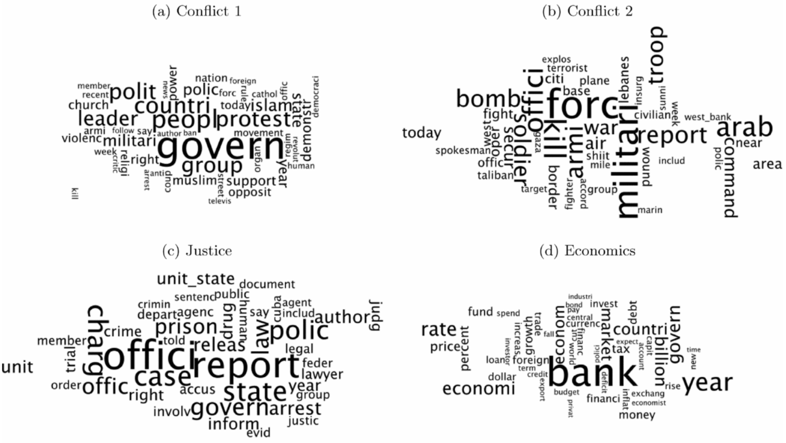
Data

- All articles on 185 countries from the New York Times (NYT), the Washington Post (WP), and the Economist for all available years since 1975. (700,000 articles in total)
- Summarize articles by their topics. (We will discuss how to do this later in the semester!)

Data

- All articles on 185 countries from the New York Times (NYT), the Washington Post (WP), and the Economist for all available years since 1975. (700,000 articles in total)
- Summarize articles by their topics. (We will discuss how to do this later in the semester!)
- Use topics to predict political violence / wars.

Topics from newspapers



Language Change Across Years

People use different languages to describe conflicts.

Both Years	Only 1995	Only 2015
forc	unit	bomb
militari	serb	american
attack	nation	group
armi	unit_nation	islam
kill	lebanes	secur
troop	defens	peopl
soldier	mile	polic
offici	gulf	citi
war	weapon	wound
report	aircraft	shiit
fight	missil	taliban
command	ship	milit
govern	plane	insurg
rebel	use	leader
guerrilla	christian	men
civilian	tank	terrorist
arm	town	violenc
border	western	northern
area	peac	capit
oper	say	sunni
offic		
base		
air		
near		
southern		
fighter		
muslim		
today		
control		
week		

We have too much data?!

$$\text{violence} = \beta_0 + \beta \cdot \text{newspaper and country features} + \epsilon$$

- Country features, such as GDP and previous conflicts.

We have too much data?!

$$\text{violence} = \beta_0 + \beta \cdot \text{newspaper and country features} + \epsilon$$

- Country features, such as GDP and previous conflicts.
- 700,000 articles, and each contain multiple topics.

We have too much data?!

$$\text{violence} = \beta_0 + \beta \cdot \text{newspaper and country features} + \epsilon$$

- Country features, such as GDP and previous conflicts.
- 700,000 articles, and each contain multiple topics.
- What should we put in the **newspaper features** list?

  Entertainment, food, sports, economy, and etc

We have too much data?!

$$\text{violence} = \beta_0 + \beta \cdot \text{newspaper and country features} + \epsilon$$

- Country features, such as GDP and previous conflicts.
- 700,000 articles, and each contain multiple topics.
- What should we put in the **newspaper features** list?

  Entertainment, food, sports, economy, and etc
- How do we decide?

Lasso can decide for us!

- Least absolute shrinkage and selection operator.
- A slightly different loss function:

Lasso can decide for us!

- Least absolute shrinkage and selection operator.
- A slightly different loss function:

$$\hat{\beta} = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta} X \right)^2 + \lambda |\widetilde{\beta}|$$

The $\frac{1}{2}$ is just a standardizing constant.

Lasso can decide for us!

- Least absolute shrinkage and selection operator.
- A slightly different loss function:

$$\hat{\beta} = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta} X \right)^2 + \lambda |\widetilde{\beta}|$$

The $\frac{1}{2}$ is just a standardizing constant.

- The added term will **shrink** the original OLS coefficient:
  - To zero if it is smaller than a certain threshold.
  - To a smaller number in absolute value if it is greater than a certain threshold.

How does Lasso work?

$$\hat{\beta} = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta} X \right)^2 + \lambda |\widetilde{\beta}|$$

When $\hat{\beta} > 0$,

$$\begin{aligned} \frac{\partial}{\partial \beta} &= \sum_{i}^{N} \left(Y_i - \widetilde{\beta} X_i \right) (-X_i) + \lambda \\ &= \sum_{i}^{N} \widetilde{\beta} X_i^2 - \sum_{i}^{N} X_i Y_i + \lambda \end{aligned}$$

How does Lasso work?

$$\hat{\beta} = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta} X \right)^2 + \lambda |\widetilde{\beta}|$$

When $\hat{\beta} > 0$,

$$\frac{\partial}{\partial \beta} = \sum_{i}^{N} \left(Y_i - \widetilde{\beta} X_i \right) (-X_i) + \lambda$$

$$= \sum_{i}^{N} \widetilde{\beta} X_i^2 - \sum_{i}^{N} X_i Y_i + \lambda$$

When $\hat{\beta} < 0$,

$$\frac{\partial}{\partial \beta} = \sum_{i}^{N} \left(Y_i - \widetilde{\beta} X_i \right) (-X_i) - \lambda$$

$$= \sum_{i}^{N} \widetilde{\beta} X_i^2 - \sum_{i}^{N} X_i Y_i - \lambda$$

When $\hat{\beta} > 0$,

$$\sum_i^N \widetilde{\beta} X_i^2 - \sum_i^N X_i Y_i + \lambda = 0$$

$$\begin{aligned}\hat{\beta} &= \frac{\sum_i^N X_i Y_i - \lambda}{\sum_i X_i^2} \\ &= \frac{A - \lambda}{B} \\ &= \left(\beta_{\textsf{OLS}} - \frac{\lambda}{B}\right) \\ &\quad \mathbf{1}\left(\beta_{\textsf{OLS}} > \frac{\lambda}{B}\right)\end{aligned}$$

When $\hat{\beta} > 0$,

$$\sum_i^N \widetilde{\beta} X_i^2 - \sum_i^N X_i Y_i + \lambda = 0$$

$$\begin{aligned}\hat{\beta} &= \frac{\sum_i^N X_i Y_i - \lambda}{\sum_i X_i^2} \\ &= \frac{A - \lambda}{B} \\ &= \left(\beta_{\mathsf{OLS}} - \frac{\lambda}{B}\right) \\ &\mathbf{1}\left(\beta_{\mathsf{OLS}} > \frac{\lambda}{B}\right)\end{aligned}$$

When $\hat{\beta} < 0$,

$$\sum_i^N \widetilde{\beta} X_i^2 - \sum_i^N X_i Y_i - \lambda = 0$$

$$\begin{aligned}\hat{\beta} &= \frac{\sum_i^N X_i Y_i + \lambda}{\sum_i X_i^2} \\ &= \frac{A + \lambda}{B} \\ &= \left(\beta_{\mathsf{OLS}} + \frac{\lambda}{B}\right) \\ &\mathbf{1}\left(\beta_{\mathsf{OLS}} < \frac{\lambda}{B}\right)\end{aligned}$$

$$\hat{\beta}_{\textsf{lasso}} = \left(\beta_{\textsf{OLS}} - \textsf{sgn}_{\beta_{\textsf{OLS}}} \cdot \frac{\lambda}{B}\right) \mathbf{1}\left(|\beta_{\textsf{OLS}}| > \frac{\lambda}{B}\right)$$

- Assume we have $\beta_{\textsf{OLS}}$ to begin with.

$$\hat{\beta}_{\text{lasso}} = \left(\beta_{\text{OLS}} - \text{sgn}_{\beta_{\text{OLS}}} \cdot \frac{\lambda}{B}\right) \mathbf{1}\left(|\beta_{\text{OLS}}| > \frac{\lambda}{B}\right)$$

- Assume we have β_{OLS} to begin with.
- Researcher can select λ.

$$\hat{\beta}_{\text{lasso}} = \left(\beta_{\text{OLS}} - \text{sgn}_{\beta_{\text{OLS}}} \cdot \frac{\lambda}{B}\right) \mathbf{1}\left(|\beta_{\text{OLS}}| > \frac{\lambda}{B}\right)$$

- Assume we have β_{OLS} to begin with.
- Researcher can select λ.
- If the original β_{OLS} has an absolute value smaller than $\frac{\lambda}{B}$, LASSO will "ditch" the variable.

$$\hat{\beta}_{\text{lasso}} = \left(\beta_{\text{OLS}} - \text{sgn}_{\beta_{\text{OLS}}} \cdot \frac{\lambda}{B}\right) \mathbf{1}\left(|\beta_{\text{OLS}}| > \frac{\lambda}{B}\right)$$

- Assume we have β_{OLS} to begin with.
- Researcher can select λ.
- If the original β_{OLS} has an absolute value smaller than $\frac{\lambda}{B}$, LASSO will "ditch" the variable.
- Otherwise, LASSO will **shrink** the value of β_{OLS} by $\frac{\lambda}{B}$.

Reading between the lines (Mueller and Rauh, 2017)

- We don't want to include too many variables.

  e.g. A topic about PBJ will add more noise (if not bias) to the estimate.

Reading between the lines (Mueller and Rauh, 2017)

- We don't want to include too many variables.

  e.g. A topic about PBJ will add more noise (if not bias) to the estimate.

- We don't want to include too few variables.

  Omitted variable bias!

Reading between the lines (Mueller and Rauh, 2017)

- We don't want to include too many variables.

  e.g. A topic about PBJ will add more noise (if not bias) to the estimate.

- We don't want to include too few variables.

  Omitted variable bias!

- It is hard for human to decide, given the amount of topics and texts.

Reading between the lines (Mueller and Rauh, 2017)

- We don't want to include too many variables.

  e.g. A topic about PBJ will add more noise (if not bias) to the estimate.

- We don't want to include too few variables.

  Omitted variable bias!

- It is hard for human to decide, given the amount of topics and texts.

- Lasso can help!

Results

Selectivity Level	Mild	Regular	Very	Mild	Regular	Very
	Civil war onset next year			Armed conflict onset next year		
	(1)					
Topic shares						
conflict1	0.0366					
	(0.0685)					
conflict2	0.256**					
	(0.104)					
justice	–0.158**					
	(0.0664)					
international relations2	–0.236**					
	(0.102)					
civic life2	–0.0869*					
	(0.0518)					
asia	–0.180**					
	(0.0803)					
sports	–0.0490					
	(0.0365)					
politics	–0.141***					
	(0.0472)					
business	–0.136**					
	(0.0549)					
economics						
Other variables						
25+ battle death	0.0699***					
	(0.0163)					
democracy score	4.81e-05					
	(0.000198)					
partial autocracy						
partial dem. with factionalism						
partial dem. w/o factionalism	0.0154					
	(0.0105)					
full democracy	0.0174*					
	(0.0102)					
4+ neighbouring conflicts	0.0247					
	(0.0396)					
child mortality rate						
ln (child mortality rate)	0.00707					
	(0.00531)					
% pop. discriminated	0.111*					
	(0.0604)					
% pop. excluded from power						
Country fixed effects	yes					
Observations	4,561					
R-squared	0.039					
Number of countries	140					
% topics in model	**56%**					

Results

Selectivity Level	Mild	Regular	Very	Mild	Regular	Very
	Civil war onset next year			Armed conflict onset next year		
	(1)	(2)				
Topic shares						
conflict1	0.0366	0.0564				
	(0.0685)	(0.0599)				
conflict2	0.256**	0.300***				
	(0.104)	(0.103)				
justice	−0.158**	−0.115*				
	(0.0664)	(0.0617)				
international relations2	−0.236**					
	(0.102)					
civic life2	−0.0869*	−0.00783				
	(0.0518)	(0.0370)				
asia	−0.180**	−0.151**				
	(0.0803)	(0.0734)				
sports	−0.0490					
	(0.0365)					
politics	−0.141***					
	(0.0472)					
business	−0.136**					
	(0.0549)					
economics						
Other variables						
25+ battle death	0.0699***	0.0713***				
	(0.0163)	(0.0164)				
democracy score	4.81e-05					
	(0.000198)					
partial autocracy						
partial dem. with factionalism						
partial dem. w/o factionalism	0.0154					
	(0.0105)					
full democracy	0.0174*					
	(0.0102)					
4+ neighbouring conflicts	0.0247					
	(0.0396)					
child mortality rate						
ln (child mortality rate)	0.00707					
	(0.00531)					
% pop. discriminated	0.111*	0.108*				
	(0.0604)	(0.0616)				
% pop. excluded from power						
Country fixed effects	yes	yes				
Observations	4,561	4,644				
R-squared	0.039	0.034				
Number of countries	140	141				
% topics in model	**56%**	**71%**				

Results

Selectivity Level	Mild	Regular	Very	Mild	Regular	Very
	Civil war onset next year			Armed conflict onset next year		
	(1)	(2)	(3)			
Topic shares						
conflict1	0.0366	0.0564				
	(0.0685)	(0.0599)				
conflict2	0.256**	0.300***	0.281***			
	(0.104)	(0.103)	(0.0961)			
justice	−0.158**	−0.115*	−0.117**			
	(0.0664)	(0.0617)	(0.0541)			
international relations2	−0.236**					
	(0.102)					
civic life2	−0.0869*	−0.00783	−0.0247			
	(0.0518)	(0.0370)	(0.0298)			
asia	−0.180**	−0.151**	−0.142**			
	(0.0803)	(0.0734)	(0.0650)			
sports	−0.0490					
	(0.0365)					
politics	−0.141***					
	(0.0472)					
business	−0.136**					
	(0.0549)					
economics						
Other variables						
25+ battle death	0.0699***	0.0713***	0.0749***			
	(0.0163)	(0.0164)	(0.0165)			
democracy score	4.81e-05					
	(0.000198)					
partial autocracy						
partial dem. with factionalism						
partial dem. w/o factionalism	0.0154					
	(0.0105)					
full democracy	0.0174*					
	(0.0102)					
4+ neighbouring conflicts	0.0247					
	(0.0396)					
child mortality rate						
ln (child mortality rate)	0.00707					
	(0.00531)					
% pop. discriminated	0.111*	0.108*				
	(0.0604)	(0.0616)				
% pop. excluded from power						
Country fixed effects	yes	yes	yes			
Observations	4,561	4,644	4,931			
R-squared	0.039	0.034	0.030			
Number of countries	140	141	143			
% topics in model	**56%**	**71%**	**80%**			

Results

Selectivity Level	Mild	Regular	Very	Mild	Regular	Very
	Civil war onset next year			Armed conflict onset next year		
	(1)	(2)	(3)	(4)	(5)	(6)
Topic shares						
conflict1	0.0366	0.0564		0.306**	0.259**	0.275***
	(0.0685)	(0.0599)		(0.121)	(0.103)	(0.0999)
conflict2	0.256**	0.300***	0.281***	0.304**		
	(0.104)	(0.103)	(0.0961)	(0.117)		
justice	−0.158**	−0.115*	−0.117**	−0.256***	−0.215***	−0.206***
	(0.0664)	(0.0617)	(0.0541)	(0.0826)	(0.0712)	(0.0705)
international relations2	−0.236**			−0.130	−0.0554	
	(0.102)			(0.0992)	(0.0909)	
civic life2	−0.0869*	−0.00783	−0.0247	−0.0196	−0.0679	
	(0.0518)	(0.0370)	(0.0298)	(0.0671)	(0.0520)	
asia	−0.180**	−0.151**	−0.142**			
	(0.0803)	(0.0734)	(0.0650)			
sports	−0.0490					
	(0.0365)					
politics	−0.141***					
	(0.0472)					
business	−0.136**					
	(0.0549)					
economics				−0.0256		
				(0.0891)		
Other variables						
25+ battle death	0.0699***	0.0713***	0.0749***			
	(0.0163)	(0.0164)	(0.0165)			
democracy score	4.81e-05					
	(0.000198)					
partial autocracy				0.0244	0.0270*	
				(0.0151)	(0.0145)	
partial dem. with factionalism				−0.00845	−0.00163	−0.00888
				(0.0124)	(0.0104)	(0.00981)
partial dem. w/o factionalism	0.0154					
	(0.0105)					
full democracy	0.0174*			0.00183	0.00442	
	(0.0102)			(0.0165)	(0.0118)	
4+ neighbouring conflicts	0.0247					
	(0.0396)					
child mortality rate				−3.86e-05		
				(0.000212)		
ln (child mortality rate)	0.00707			0.00376		
	(0.00531)			(0.00852)		
% pop. discriminated	0.111*	0.108*				
	(0.0604)	(0.0616)				
% pop. excluded from power				−0.0488		
				(0.0442)		
Country fixed effects	yes	yes	yes	yes	yes	yes
Observations	4,561	4,644	4,931	3,991	4,226	4,226
R-squared	0.039	0.034	0.030	0.012	0.008	0.006
Number of countries	140	141	143	138	139	139
% topics in model	**56%**	**71%**	**80%**	**50%**	**57%**	**67%**

A Natural Experiment

- Parental leave policy leads to gender stereotypes.

A Natural Experiment

- Parental leave policy leads to gender stereotypes.
- Having a father's leave policy could reduce sexist attitudes.

A Natural Experiment

- Parental leave policy leads to gender stereotypes.
- Having a father's leave policy could reduce sexist attitudes.
- Estonia prolonged father's leave after July 1st 2020.

A Natural Experiment

- Parental leave policy leads to gender stereotypes.
- Having a father's leave policy could reduce sexist attitudes.
- Estonia prolonged father's leave after July 1st 2020.
- Parents who gave birth on June 30th and July 1st are almost **randomly** assigned into treatment and control group.

  Nature (or God) determines whether the baby is born before or after July 1st, if the mother's due date is close to that date.

What are the useful variables to control for?

$$\text{Sexist attitudes} = \beta_0 + \beta_1 \text{treatment} + \beta \text{socio-economic covariates} + \epsilon_i$$

- What should we include in the socio-economic covariates?

  Age, educaton, income, marriage status, employment, race and ethnicity and etc.

What are the useful variables to control for?

$$\text{Sexist attitudes} = \beta_0 + \beta_1\text{treatment} + \beta\text{socio-economic covariates} + \epsilon_i$$

- What should we include in the socio-economic covariates?

  Age, educaton, income, marriage status, employment, race and ethnicity and etc.
- Lasso can help us!

What are the useful variables to control for?

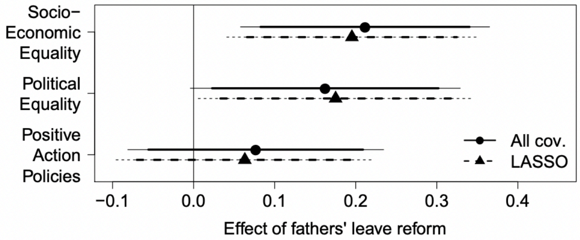
$$\text{Sexist attitudes} = \beta_0 + \beta_1 \text{treatment} + \beta \text{socio-economic covariates} + \epsilon_i$$

- What should we include in the socio-economic covariates?

  Age, educaton, income, marriage status, employment, race and ethnicity and etc.
- Lasso can help us!
- Authors show results both with and without Lasso selection.

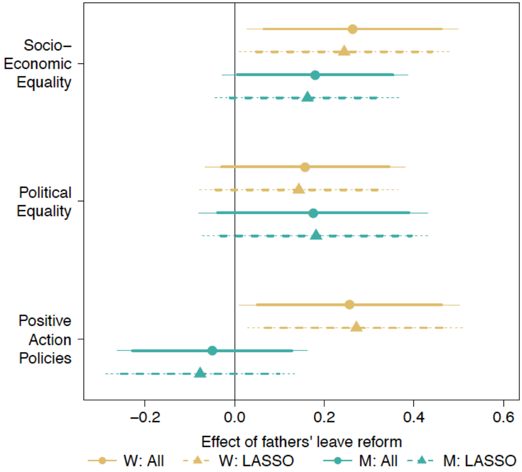
Increase support of gender equality after treatment

Figure 1: Effect of fathers' leave reform on gender-equal attitudes, Study 1



Mothers and fathers responded similarly

Figure 2: Effect of fathers' leave reform on gender-equal attitudes for mothers and fathers, Study 1



Results

- Compared to unaffected parents, increased support of equality.
- Mothers and fathers responded similarly, even though the policy only prolonged father's leave.
- Slight difference between all covariates vs Lasso selected covariates.

Details and Critique

- λ is always greater or equal to zero.

Details and Critique

- λ is always greater or equal to zero.
- Researchers can decide how selective LASSO model is.

Details and Critique

- λ is always greater or equal to zero.
- Researchers can decide how selective LASSO model is.
  - Parameter tuning / cross validation.

Details and Critique

- λ is always greater or equal to zero.
- Researchers can decide how selective LASSO model is.
  - Parameter tuning / cross validation.
- LASSO improves prediction performance, at the cost of a biased coefficient.

Wald Estimator

Under one sided compliance, we can simplify Wald Estimator as:

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(T_i|Z_i = 1)}$$

This is true because $E(T_i|Z_i = 0)$ is zero for both compliers and never-takers.

Let's then take a closer look at the numerator.

$$
\begin{aligned}
&E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) \\
&= E(Y_i|Z_i = 1) - \underbrace{E(Y_i(0)|Z_i = 0)}_{\text{one-sided compliance}} \\
&= E(Y_i(1)|Z_i = 1, T_i = 1)P(T_i = 1|Z_i = 1) \\
&\quad + E(Y_i(0)|Z_i = 1, T_i = 0)P(T_i = 0|Z_i = 1) - E(Y_i(0)|Z_i = 0) \\
&= E(Y_i(1)|Z_i = 1, T_i = 1)P(T_i = 1|Z_i = 1) \\
&\quad - E(Y_i(0)|Z_i = 1, T_i = 1)P(T_i = 1|Z_i = 1) \\
&+ \underbrace{E(Y_i(0)|Z_i = 1, T_i = 1)P(T_i = 1|Z_i = 1) + E(Y_i(0)|Z_i = 1, T_i = 0)P(T_i = 0|Z_i = 1)}_{E(Y_i(0)|Z_i=1)} \\
&= E(Y_i(1)|Z_i = 1, T_i = 1)P(T_i = 1|Z_i = 1) \\
&\quad - E(Y_i(0)|Z_i = 1, T_i = 1)P(T_i = 1|Z_i = 1) \\
&\quad + \underbrace{E(Y_i(0)|Z_i = 1) - E(Y_i(0)|Z_i = 0)}_{=0 \text{ due to randomization}} \\
&= E(Y_i(1) - Y_i(0)|Z_i = 1, T_i = 1)P(T_i = 1|Z_i = 1) \\
&= E(Y_i(1) - Y_i(0)|T_i = 1)E(T_i|Z_i = 1) \quad \text{exclusion restriction}
\end{aligned}
$$

Thus we have:

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(T_i|Z_i = 1)} = E(Y_i(1) - Y_i(0)|T_i = 1)$$