Comparing Performance of Lasso, Group Lasso, and Linear Regression with Categorical Predictors

Yihuan Huang, Amanda K. Montoya

University of California, Los Angeles

UCLA Quantitative Research Collaboratory

2019 Symposium on Data Science and Statistics

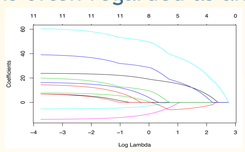
May 31, 2019 Seattle, WA

Introduction

Machine learning is used frequently to train models and predict outcomes in different scientific areas.

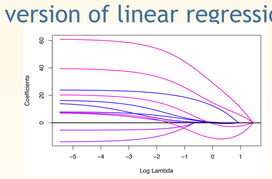
- ***Lasso regression*** performs variable selection and regularization. It is often regarded as an advanced version of linear regression[1].

Lasso



$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}}(\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|)$$

Group Lasso



$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}}(\|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_{I_g}\|_2)$$

- ***Group lasso*** is an alternative to lasso to align with properties from linear regression, for models with categorical predictors.

Researchers, especially in social science field, primarily focus on similarities between linear regression and lasso, but pay little attention to their different properties, particularly involving categorical predictors.

We aim to show that linear regression, lasso, and group lasso have distinct pros and cons and should be treated accordingly.

- For example, ***coding strategies*** are used to include categorical predictors in linear regression. In this project, we examine their performance for lasso and group lasso.

1	0	0
0	1	0
0	0	1
0	0	0

Dummy

1	0	0
0	1	0
0	0	1
-1	-1	-1

Effect

-3/4	0	0
1/4	-2/3	0
1/4	1/3	-1/2
1/4	1/3	1/2

Helmert

Study I

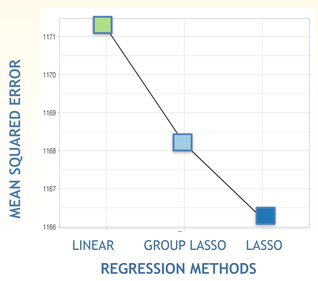
RESEARCH QUESTION:

Across different regression methods
how do predicted group means change?

REGRESSION METHODS
- Linear regression
- Lasso regression
- Group lasso regression

GRAPH

Sample size = 3000
Training data size = 1000
Testing data size = 2000



WAGE DATASET
- Outcome variable: wage
- Predictor:
  - 7 categorical variables
  - 1 continuous variable
- Dummy coding for categorical variables
- Every first group as reference group

Study II(1)

RESEARCH QUESTION:

Across different reference groups using lasso
how do coefficients change?

MEASURES
- Use lasso and dummy coding strategy for categorical variables
- Build five different models corresponding to different choices of reference groups with the category "marital status"

DIFFERENCES IN VARIABLE SELECTION AND COEFFICIENTS

		Reference Groups				
		Single	Married	Widowed	Divorced	Separated
Coefficients	Intercept	96.413	122	98	96	96
	Single	.	-25.65	0	0	0
	Married	24.996	.	23.369	25.168	25.535
	Widowed	10.27	-4.07	.	11.746	13.95
	Divorced	0	-25.14	0	.	0
	Separated	-0.462	-28	0	-0.8767	.

Study II(2)

RESEARCH QUESTION:

Across different methods and coding strategies
how does variable selection and model fit change?

DIFFERENCES IN VARIABLE SELECTION

Variable not selected in model

Variable selected in model

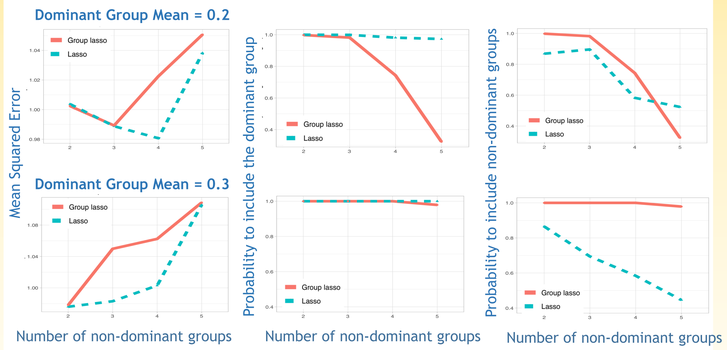
Lasso Regression			Group Lasso Regression		
Dummy	Effect	Helmert	Dummy	Effect	Helmert
	Intercept			Intercept	
	Married			Married	
	Widowed			Widowed	
	Divorced			Divorced	
	Separated			Separated	
	Black			Black	
	Asian			Asian	
	Other			Other	
	High School Grad			High School Grad	
	Some College			Some College	
	College Grad			College Grad	
	Advanced Degree			Advanced Degree	
	Job Class			Job Class	
	Health			Health	
	Health Insurance			Health Insurance	
	Age			Age	
MSE 1172.76	1166.91	1168.30	1169.16	1167.34	1166.66

Study III

RESEARCH QUESTION

Will group lasso cause **over fitting** issues?

Monte Carlo Simulation



METHODS
- Apply lasso and group lasso to simulated dataset
- Set dominant group mean = 0.2 or 0.3
- Calculate probabilities that
  - Models include the dominant group
  - Models include non-dominant groups

MODEL
- One categorical variable
  - A dominant group
  - Several non-dominant groups

Discussion

CONCLUSION + FUTURE DIRECTIONS

- ***Prediction accuracy***
  - Lasso>Group Lasso>Linear Regression
- ***Lasso*** differs from linear regression
  - **Heavily depends** on coding strategies.
- ***Group lasso***
  - Always perform **same** variable selection
  - Prediction accuracy depends on choices of coding strategies.
  - Cause **over fitting issue** when there is a dominant group within a categorical variable.
    - More likely to include non-predictive groups than lasso, which decreases the prediction accuracy of the model.

WHAT'S NEXT?

Design a machine learning algorithm that both the variable selection and prediction accuracy of the model are independent of the choices of coding strategies

References

1. Patel, P. C. (2018) The Great Recession and allostatic load in the United States. *International Journal of Stress Management, 10.*