Performance of LASSO when One or More Covariates are Missing Not at Random

M.Sc. Project

Rajesh Majumder

Department of Statistics,
West Bengal State University

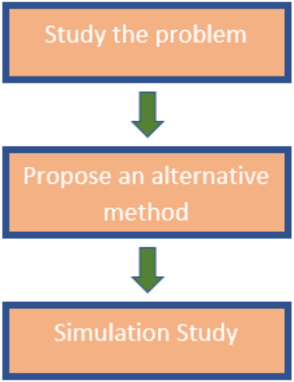
August 23rd, 2021

Introduction

- In any survey method, missing data is a common problem, and in regression when there are huge number of covariates, there must be a linear dependency among them.
- As, Least Absolute Selection and Shrinkage Operator (LASSO) is a variable selection and also a estimation technique, in this project, we are trying to see how LASSO will perform the variable selection task when one or more linearly dependent covariates are affected by the not at random missing mechanism.

Overview

Figure: *OVERVIEW OF OUR WORK*



What is LASSO ?

- Least Absolute Selection and Shrinkage Operator (LASSO) is a new technique, proposed by R.Tibshirani (1996) based on Breiman's non-negative garrote.
- The LASSO estimate is defined as

$$\widehat{\beta}^{lasso} = \underset{\beta \in \mathbb{R}^{\prime}}{arg\ min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \right\}$$

$$\widehat{\beta}^{lasso} = \underset{\beta \in \mathbb{R}^{\prime}}{arg\ min} \left\{ \underbrace{\parallel y - X\beta \parallel_2^2}_{Loss} + \lambda \underbrace{\parallel \beta \parallel_1}_{Penalty} \right\}$$

- So, this is a l_1- penalized ordinary least square technique.
- $\lambda \geq 0$ penalty parameter.
- Penalty deals collinearity.
- This is a high dimensional technique, also applicable in low dimension.

What is LASSO ?

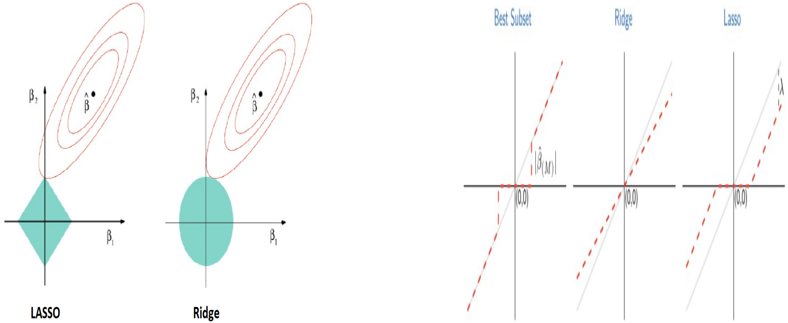


Figure: **Left:** *Ridge V/S. LASSO* / **Right:** *Best Subset Selection V/S. Ridge V/S. LASSO*

Computation of LASSO :

- It is a convex function but not differentiable.
- To solve this, there are different techniques : Coordinate descent method, Gradient descent method, LARS algorithm etc.
- In this progect we are using LARS method.
- LARS = Least-angle regression + Forward Stagewise Algorithm.
- In R, use *lars()* function under *lars* package .
- To choose λ in general, we do 10-fold Cross Validation technique.
- And , to get the estimated values of β, we use 1-standard error rule.

Missing Data and Missing Not at Random :

- Missing data is a common problem in any survey problem.
- Missingness causes loss of information and gives wrong inference.
- There are two types of missing data, ignorable & non-ignorable.
- Rubin (1976) has proposed different missing mechanism : Missing Completely at Random(MCAR), Missing at Random(MAR) & Missing Not at Random(MNAR).
- **MNAR** :

$$P_{R|Z}\left(R_i = r_i \mid Z_i\right) = P_R\left(R_i = r_i \mid Z_{(\bar{r})}\right)$$

where $R = 0 \; or \; 1 \;\; if \;\; Z \; is \; missing$ = missing Indicator varable and $r = (0 \; or \; 1)$ observed value of R. $i = 1, 2, ..., n$

Problem of LASSO Estimation under Missing Covariate

- Consider the regression model : $Y = \beta_1 X_1 + ... + \beta_p X_p + \epsilon$.
- WLG, say X_1 is missing under MNAR mechanism.
- We define, $X_i = (X_{1i}, ..., X_{pi})^T$, $X_i^* = (X_{2i}, ..., X_{pi})^T$,$R_i = 1$ *or* 0 if X_{1i} is missing or not.
- Also define,
  - $\mathcal{O} = \{i : R_i = 1;\ \forall i = 1, 2, ..., n\}$ = Set of completely observed units.
  - $\mathcal{M} = \{i : R_i = 0;\ \forall i = 1, 2, ..., n\}$ = Set of units for which X_{1i} missing.
- Now, due to linear dependency, we are assuming, X_{1i} depends on $\widetilde{X}_i^{p_1 \times 1}$, where $p_1 \geq 0$. So, $X_i = \left[X_{1i} \vdots \widetilde{X}_i^{p_1 \times 1} \vdots \widetilde{\widetilde{X}}_i^{p_2 \times 1}\right]$, $p_1 \ll p_2,\ p_1 + p_2 = p - 1\ \&\ p_2 > 1$ and $X_i = \left(X_{1i}, \widetilde{X}_i, \widetilde{\widetilde{X}}_i\right)^T$,$X_i^* = \left(\widetilde{X}_i, \widetilde{\widetilde{X}}_i\right)^T$

Problem of LASSO Estimation under Missing Covariate

- And,
  $P\left(R_i = 1 \mid X_i, Y_i\right) = P\left(R_i = 1 \mid X_{1i}, \widetilde{X}_i\right) = \pi\left(X_{1i}, \widetilde{X}_i\right)$ (say)
- Now, since LASSO works is a high dimensional set up, it performs well on completely observed data, but, due to MNAR, the reduced data set cannot be viewed as a random sample from the target population (Y, X).
- Note that,

$$\begin{aligned} f\left(y_i, x_i \mid R_i = 1\right) = & \frac{f\left(R_i = 1 \mid y_i, x_i\right) f\left(y_i, x_i\right)}{f\left(R_i = 1\right)} \\ = & W\left(x_{1i}, \widetilde{x}_i\right) \times f\left(y_i, x_i\right) \end{aligned}$$

  where $W\left(x_{1i}, \widetilde{x}_i\right) \neq 1 \ \forall \left(x_{1i}, \widetilde{x}_i\right)$.
- $\Rightarrow$sample distribution of the data from completely samples differ from the population distribution.
- $\Rightarrow\{(Y_i, X_i) : i \in \mathcal{O}\}$ is not a sufficient for estimate the target population given by the linear regression equation.

Proposed Method

- Assuming that, selection probability $\pi_i \left(= \pi(x_{1i}, \widetilde{x}_i) > 0\right)' s$ are known for the set $\mathcal{O}$; we can modify the LASSO estimator by penalizing inverse of the selection probability weighted error sum of squares.
- This is called Inverse Probability Weighted (IPW) approach in missing literature.
- This is similar as Horvitz-Thompsom estimator in survey sampling.
- So, now, the unbiased estimator of finite population total

$$S = \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2$$

is

$$S_{\mathcal{O}} = \sum_{i \in \mathcal{O}} \pi_i^{-1} \left(Y_i - X_i^T \beta\right)^2$$

Proposed Method

- So, we can use this estimate, in LASSO model to overcome the problem on variable selection under covariate missing situation.
- We are naming this as, IPW-LASSO.
- But, here is a problem, i,e., we do not know the original π_i values.
- So, we have to estimate $\pi_i's$ $\forall i \in \mathcal{O}$.
- Now, $R_i \sim Bernoulli(\pi_i)$.
- As, $X_1^{n\times 1}$ is highly correlated with $(X_2, ...X_p)$, we can use the Logistic regression technique s.t.,

$$
\begin{aligned}
ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha_1 X_{1i} + \alpha_2^T \widetilde{X}_i \\
&\approx \theta_1 X_{2i} + ... + \theta_{p-1} X_{pi} \\
&= \theta^T X_i^*
\end{aligned}
$$

Proposed Method

- Hence, for the i^{th}unit,

$$\begin{aligned} P\left(R_i = 1 \mid X_{1i}, \widetilde{X}_i\right) \approx P\left(R_i = 1 \mid X_i^*\right) &= \frac{exp\left(\theta^T X_i^*\right)}{1 + exp\left(\theta^T X_i^*\right)} \\ &= \pi_i^*\left(X_i^*\right) \end{aligned}$$

- So, this is a MAR mechanism.
- so, to estimate $\pi_i^*\left(X_i^*\right) \pi_i^*\left(X_i^*\right)\left(X_i^*\right)$, we are using Logistic Regression technique (MLE).
- Using $\widehat{\pi_i^*}^{MLE}$, the IPW-LASSO mode is :$\widehat{\beta}^{IPW-LASSO} =$ $\underset{\beta \in \mathbb{R}^{\prime}}{arg\ min} \left\{ \sum_{i=1}^{n} \left(\widehat{\pi_i^*}^{MLE}\right)^{-1} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \right\}$
- But this will give a wrong variable selection result.

Proposed Method

- To overcome this,we are using two-step LASSO techniques.
- Estimating π_i^*, by l_1-penalized Logistic regression method.
- Obtain $\widehat{\pi_i^*}^{LASSO}$.
- Using $\widehat{\pi_i^*}^{LASSO}$, the IPW-LASSO mode is :$\widehat{\beta}^{IPW-LASSO} =$
$$\underset{\beta\in\mathbb{R}^p}{arg\ min}\left\{\sum_{i=1}^{n}\left(\widehat{\pi_i^*}^{MLE}\right)^{-1}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\mid\beta_j\mid\right\}$$

Simulation Study

Simulation Steps

- The regression model is
  $Y_i = \beta_1 X_{1i} + ... + \beta_{12} X_{12i} + \epsilon_i \ \forall i = 1(1)200$
- Generating 10 random variables $\{X_1, ..., X_{10}\}$ from multivariate normal distribution with $\mu_i = 0$ and variance $\sigma_j^2 = 1 \ \forall j = 1(1)10$
- And the $\sum^{10\times10}$ matrix with diagonal elements are $diag(1,1,.....1)$ and off-diagonals are in the form

$$\rho_{j,k} = \rho^{|j-k|} \quad \forall j,k = 1(1)10 \ j < k$$

- Generating two variables (X_{11}, X_{12}) from Exponential distribution with mean$(\theta) = 2$ and from Uniform distribution $U(0,1)$ respectively.
- Generating the error variable $\epsilon \sim N_{200}(0, \sigma^2 I_{200})$. And based on these, generating $Y^{200\times1}$.
- WLG, $X_5$is missing and X_5 is depending on X_4.

Simulation Steps

- Then, generating missing data and applying five different types LASSO on original and complete observed data sets.

LASSO on original Data set
(when all known)

↓

LASSO on Complete Data Set
(Removing All Missing Observations)

↓

IPW-LASSO on Complete Data Set using
known(actual) $\pi_i^* (X_i^*)$

↓

IPW-LASSO on Complete Data Set using
Estimated(MLE) $\pi_i^* (X_i^*)$

↓

IPW-LASSO on Complete Data Set using
Estimated(Logistic Lasso) $\pi_i^* (X_i^*)$

Simulation Steps

- Repeat this 500 times.
- Note that, here, $(\beta_1, \beta_2, \beta_5, \beta_6, \beta_8) = (1, 1, 1.5, 1, 1)$ and rest of the $\beta's$ are zero.
- $(\sigma^2, \rho) = (2, 0.5)$
- and the Logistic regression coefficients are $(\theta_3, \theta_4) = (0.1, 0.1)$, corresponding to the variables $X_4$and X_5 and rest of the others are zero.
- And, our concern is that, average 45% missingness for X_5.

True Beta	LASSO On Original Simulated Data set	LASSO on Completely observed data	IPW-LASSO with known observed sample probability	IPW-LASSO with estimated (MLE) observed sample probability	IPW-LASSO with estimated (Logistic LASSO) observed sample probability
1	0.7408	0.6433	0.6674	0.6674	0.6708
1	0.7701	0.7063	0.7063	0.7021	0.7043
0	0	0.0262	0.0241	0.0245	0.0239
0	0	0.0271	0	0.0265	0.0032
1.5	1.3502	1.1985	1.2932	1.189	1.1968
1	0.8296	0.7658	0.8275	0.7647	0.7932
0	0	0.0493	0.0484	0.0517	0.0474
1	0.8196	0.5678	0.6902	0.5515	0.5649
0	0	0.0089	0	0.0088	0
0	0	0	0	0.0079	0
0	0	0	0	0	0
0	0	0	0	0.0024	0

Figure: Average Simulated $\hat{\beta}$ estimates for Different methods

Simulation Result

Methods	Average False Selection (AFS)
LASSO On Original Simulated Data set	0.0578
LASSO on Completely observed data	0.0762
IPW-LASSO with known observed sample probability	0.0737
IPW-LASSO with estimated (MLE) observed sample probability	0.805
IPW-LASSO with estimated (Logistic LASSO) observed sample probability	0.747

Figure: Average False selection table

Simulation Result

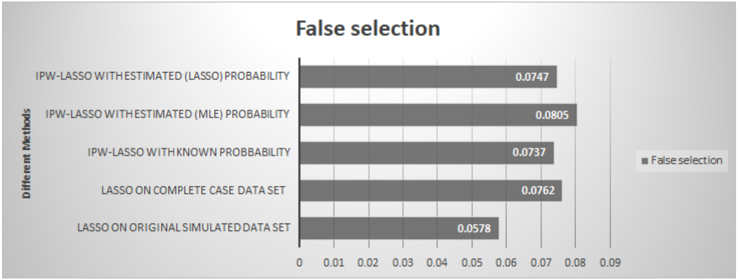


Figure: Average False selection Column diagram

Thank You, . . .