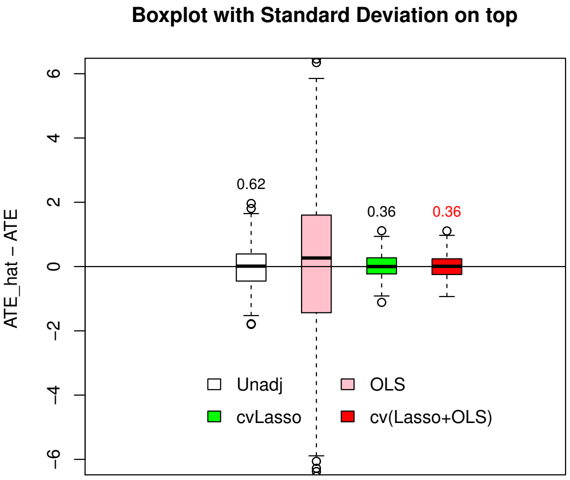
Adjusting treatment effect estimates in randomized experiments with the Lasso

Jasjeet S. Sekhon

Departments of Political Science and Statistics
University of California, Berkeley
https://sekhon.berkeley.edu

Joint work with Adam Bloniarz, Hanzhong Liu, Bin Yu, and Cunhui Zhang

A simulation study based on experimental data with a large number (p) of covariates relative to sample size n



n=59*2, p=59, p_A=0.5

Our plan

- Theoretical study of Lasso under Neyman-Rubin model to gain insights into when Lasso works as an adjustment method.
- Simulation and real data experiments to argument theoretical study about regularization parameter selection and compare Lasso and its variants.
- Estimation model is not being assumed. Useful framework to study other ML methods: random forests, SVM, deep learning, etc.

Related work

- Regression adjustment for fix p under Neyman-Rubin:
  - Freedman DA 2008;
  - Lin W 2013;
- Regression adjustment for $p > n$ under regression model:
  - Belloni A, Chernozhukov V, Hansen C 2013;
  - Belloni A, Chernozhukov V, Fernandez-Val I, Hansen C 2013;
  - Tian L, Alizadeh A, Gentles A, Tibshirani R 2014;
  - Rosenblum M, Liu H, En-Hsu Y 2014;

Other Related Work

- Bowers, Panagopoulos, and Fredrickson 2013; Bowers 2014; Bowers, Fredrickson, Hansen 2015
- Imai and Ratkovic 2013: `findit`
- Ratkovic and Tingley 2015: `sparsereg`
- Grimmer, Messing, Westwood 2014
- Athey and Imbens 2015; Wager and Athey 2016

Neyman-Rubin model

SUTVA (Rubin, 1980)
(Stable Unit Treatment Value Assumption)

- No interference
- Only a single version of each treatment level

Under SUTVA

- a_i, b_i: potential outcomes for unit i under **treatment** and **control**
- The parameter we are trying to estimate in this work is the **A**verage **T**reatment **E**ffect

$$ATE = \frac{1}{n}\sum_{i=1}^{n} a_i - \frac{1}{n}\sum_{i=1}^{n} b_i$$

Randomized experiment

- Randomness comes from treatment assignment
- T_i: random indicator of treatment for unit i
- A: set of treated units (random)

$$A = \{i, T_i = 1\}$$

- B: set of control units (random)

$$B = \{i, T_i = 0\}$$

- y_i: observed outcome

$$y_i = a_i T_i + b_i(1 - T_i), \quad i = 1, ..., n$$

Notations

- n, n_A, n_B: number of treated and control units

$$p_A = n_A/n; \; p_B = n_B/n$$

- Average on the population/treated/control

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i, \;\; \bar{a}_A = \frac{1}{n_A}\sum_{i \in A} a_i$$

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i, \;\; \bar{b}_B = \frac{1}{n_B}\sum_{i \in B} b_i$$

Simple estimator

- Understanding the **assignment mechanism** is crucial for causal inference
- If assignment is completely randomized, the ATE can be estimated without bias using the simple difference in means:

$$\widehat{ATE}_{\text{unadj}} = \bar{a}_A - \bar{b}_B$$

Regression adjustment

- This leads us to consider estimators of the form

$$\widehat{ATE} = \left[\bar{a}_A - (\bar{x}_A - \bar{x})^T \beta^{(a)}\right] - \left[\bar{b}_B - (\bar{x}_B - \bar{x})^T \beta^{(b)}\right]$$

- Regression adjustment with interaction (Lin W 2013[1])

$$y_i \sim T_i,\ x_i,\ T_i(x_i - \bar{x})$$

Minimizing

$$\sum_{i=1}^{n} \left\{ y_i - \tau_a T_i - \tau_b (1 - T_i) - T_i (x_i - \bar{x})^T \beta^{(a)} - (1 - T_i)(x_i - \bar{x})^T \beta^{(b)} \right\}^2$$

$$\widehat{ATE}_{\text{OLS}} = \left[\bar{a}_A - (\bar{x}_A - \bar{x})^T \hat{\beta}^{(a)}\right] - \left[\bar{b}_B - (\bar{x}_B - \bar{x})^T \hat{\beta}^{(b)}\right]$$

[1] Lin W (2013). Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. The Annals of Applied Statistics 7:295-318.

Regression adjustment

- Benefits of regression adjustment:

  Under regularity conditions, $\widehat{ATE}_{\text{OLS}}$ is asymptotic normal with asymptotic variance **no larger than** that of the $\widehat{ATE}_{\text{unadj}}$ (Lin W 2013)

- Question: what if $p > n$?
  - Observe many covariates
  - Main effects + interactions
  - Polynomial or splines

Regression adjustment using Lasso

- Sparsity: not all the covariates are relevant
- Lasso (Tibshirani 1996), minimizing

$$\frac{1}{2n}\sum_{i=1}^{n}\left\{y_i-\tau_a T_i-\tau_b(1-T_i)-T_i(x_i-\bar{x})^T\beta^{(a)}-(1-T_i)(x_i-\bar{x})^T\beta^{(b)}\right\}^2$$

$$+\lambda_a||\beta^{(a)}||_1+\lambda_b||\beta^{(a)}||_1$$

- Equivalent to (similarly for control group):

$$\hat{\beta}^{(a)}=\underset{\beta}{\operatorname{argmin}}\,\frac{1}{2n_A}\sum_{i\in A}\left\{a_i-\bar{a}_A-(x_i-\bar{x}_A)^T\beta\right\}^2+\lambda_a||\beta||_1$$

Regression adjustment using Lasso

- Define Lasso adjusted ATE estimator

$$\widehat{ATE}_{\text{Lasso}} = \left[\bar{a}_A - (\bar{x}_A - \bar{x})^T \hat{\beta}^{(a)}\right] - \left[\bar{b}_B - (\bar{x}_B - \bar{x})^T \hat{\beta}^{(b)}\right]$$

- We are interested in
  - Sufficient conditions for Lasso adjustment to work: Neyman-Rubin allows consideration of transformation before adjustment.
  - asymptotic normality of the adjusted effect estimate
  - estimate of the asymptotic variance

Assumptions

Assume there exist

- $S^{(a)} \subseteq \{1, 2, ..., p\}$: relevant covariates set for treatment, and
- $\beta^{(a)}$: projection coefficients.
- Then we can decompose potential outcome as follows:

$$a_i = \bar{a} + (\bar{x}_i - \bar{x})^T \beta^{(a)} + e_i^{(a)}$$

- Similar for control, then define $S = S^{(a)} \bigcup S^{(b)}$
- All the quantities above are fixed

Assumptions, similar to Freedman (2008); Lin (2013)

- Condition 1: Stability of treatment assignment probability

$$n_A/n \to p_A,\ n_B/n \to p_B,\ \text{as } n \to \infty,$$

for some $p_A, p_B \in (0,1)$

- Condition 2: The centered moment conditions

$$n^{-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^4 \le L,\ \forall j$$

$$n^{-1} \sum_{i=1}^{n} (e_i^{(a)})^4 \le L;\ \ n^{-1} \sum_{i=1}^{n} (e_i^{(b)})^4 \le L$$

- Condition 3: The means $n^{-1} \sum_{i=1}^{n} (e_i^{(a)})^2$, $n^{-1} \sum_{i=1}^{n} (e_i^{(b)})^2$ and $n^{-1} \sum_{i=1}^{n} e_i^{(a)} e_i^{(b)}$ converge to finite limits.

Two quantities needed for high-dim case

- Sparsity measures (number of nonzero coefficients)

$$s = \left|\{j : \beta_j^{(a)} \neq 0 \ \text{ or } \ \beta_j^{(b)} \neq 0\}\right|$$

- Maximum covariance

$$\delta_n = \max_{\omega=a,b} \left\{ \max_j \left| \frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right) \left(e_i^{(\omega)} - \bar{e}^{(\omega)}\right) \right| \right\}$$

Further assumptions for consistency of Lasso

- Condition 4: Decay and scaling

$$\delta_n = o\left(\frac{1}{s\sqrt{\log p}}\right); \quad (s\log p)/\sqrt{n} = o(1)$$

- Condition 5: Cone invertibility factor

$$\|h_S\|_1 \leq Cs\|\hat{\Sigma}h\|_\infty, \ \forall h \in \mathcal{C} = \{h : \|h_{S^c}\|_1 \leq \xi\|h_S\|_1\}$$

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

- Condition 6: Tuning parameter

$$\lambda_a \in (\frac{1}{\eta}, M] \times \left(\frac{11\sqrt{L}}{3p'_A}\sqrt{\frac{\log p}{n}} + \delta_n\right)$$

$$\lambda_b \in (\frac{1}{\eta}, M] \times \left(\frac{11\sqrt{L}}{3p'_B}\sqrt{\frac{\log p}{n}} + \delta_n\right)$$

Asymptotic Normality

Theorem 1

Assume conditions 1 - 6 hold. Then

$$\sqrt{n}\left(\widehat{ATE}_{\text{Lasso}} - ATE\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right),$$

$$\sigma^2 = \lim_{n\to\infty}\left[\frac{1-p_A}{p_A}\sigma^2_{e^{(a)}} + \frac{p_A}{1-p_A}\sigma^2_{e^{(b)}} + 2\sigma_{e^{(a)}e^{(b)}}\right]$$

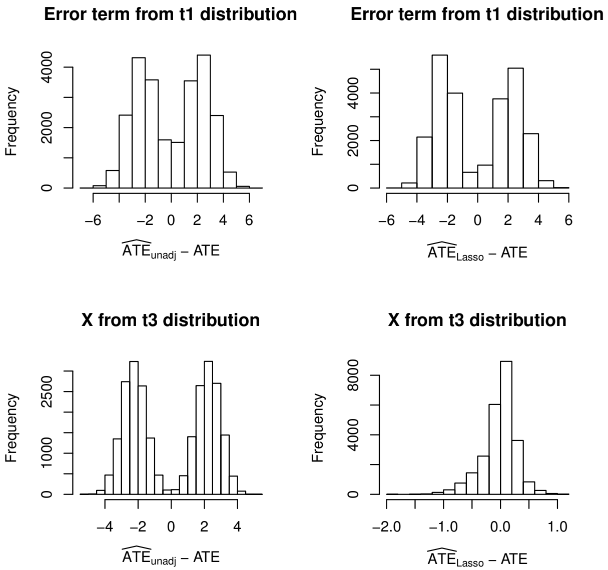
which is no greater than the asymptotic variance of the $\sqrt{n}\left(\widehat{ATE}_{\text{unadj}} - ATE\right)$*. The difference is* $\frac{1}{p_A(1-p_A)}\Delta$*.*

$$\Delta = -\lim_{n\to\infty}\|X\beta_E\|_2^2 \le 0, \quad \beta_E = (1-p_A)\beta^{(a)} + p_A\beta^{(b)}$$

Our conditions are telling

- In simulation studies, we see that some moment conditions are necessary. That is, we find that the distribution of the Lasso adjusted estimator can be non-normal when these conditions do not hold.
- Nevertheless, in our simulation studies, Lasso adjusted estimator still has a smaller MSE than the unadjusted estimator

When moment conditions fail



Conservative variance estimate

- As stated in Theorem 1, asymptotic variance

$$\sigma^2 = \lim_{n\to\infty}\left[\frac{1-p_A}{p_A}\sigma^2_{e^{(a)}} + \frac{p_A}{1-p_A}\sigma^2_{e^{(b)}} + 2\sigma_{e^{(a)}e^{(b)}}\right]$$

- Let

$$\hat{\sigma}^2_{e^{(a)}} = \frac{1}{n_A - df^{(a)}}\sum_{i\in A}\left\{a_i - \bar{a}_A - (x_i - \bar{x}_A)^T\hat{\beta}^{(a)}\right\}^2$$

$$df^{(a)} = \hat{s}^{(a)} + 1 = ||\hat{\beta}^{(a)}||_0 + 1$$

- Define

$$\hat{\sigma}^2 = \frac{n}{n_A}\hat{\sigma}^2_{e^{(a)}} + \frac{n}{n_B}\hat{\sigma}^2_{e^{(b)}}$$

- We show that: $\hat{\sigma}^2$ is asymptotically conservative estimate of σ^2

PAC-man data: results

- OLS is computed by using only 59 main effect
- PAC has no significant average treatment effect
- regression-based methods provide 20% shorter intervals
- cv(Lasso) selects 24, 8 covariates for treatment and control respectively
- cv(Lasso+OLS) selects 4, 5 covariates for treatment and control respectively;

PAC-man data: simulation

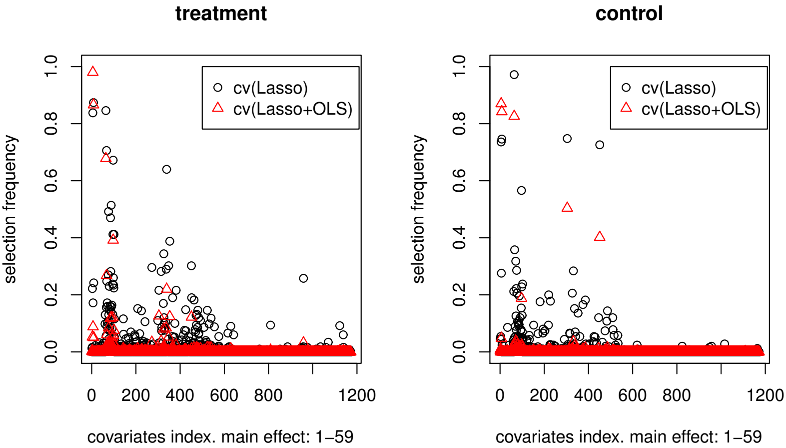
Matched on 59 main effects (ATE=-0.29), conduct 500 randomized experiments

Table 1: Results for the PAC-based simulations

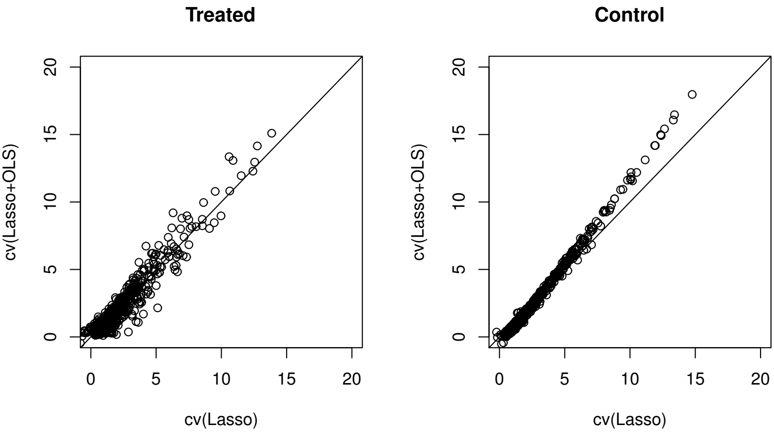
	Bias	SD	$\sqrt{MSE}$	Coverage	Length
Unadj	0.001	0.20	0.20	99%	1.06
OLS	0.002	0.18	0.18	99%	0.95
cv(Lasso)	0.001	0.17	0.17	99%	**0.94**
cv(Lasso+OLS)	**0.000**	**0.17**	**0.17**	99%	0.95

	model size	
	treatment	control
Unadj	0	0
OLS	59	59
cv(Lasso)	25	15
cv(Lasso+OLS)	6	4

Selection stability comparison: cv(Lasso) and cv(Lasso+OLS)



Adjustment value comparison: cv(Lasso) and cv(Lasso+OLS)



Bound (I): Massart concentration inequality

Proposition 1 (Massart concentration inequality for sampling without replacement)

Let $\{z_i, i = 1, ..., n\}$ be a finite population of real numbers. Let $A \subset \{i, \ldots, n\}$ be a subset of deterministic size $|A| = n_A$ that is selected randomly without replacement. Define $p_A = n_A/n,\ \sigma^2 = n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$. Then, for any $t > 0$,

$$P\left(\bar{z}_A - \bar{z} \geq t\right) \leq \exp\left\{-\frac{p_A n_A t^2}{(1+\tau)^2\sigma^2}\right\},$$

with $\tau = \min\left\{1/70, (3p_A)^2/70, (3 - 3p_A)^2/70\right\}$.

Summary of results

- Theoretical analysis of Lasso under Neyman-Rubin model, pointing to importance of moment conditions for covariates and error terms, and also cone invertibility for Lasso to work.
- Recommendation of cv(Lasso+OLS): much fewer covariates selected when compared with cv(Lasso) and with similar coverage and confidence interval length.
- Estimating Heterogeneous treatment effects?

PAC

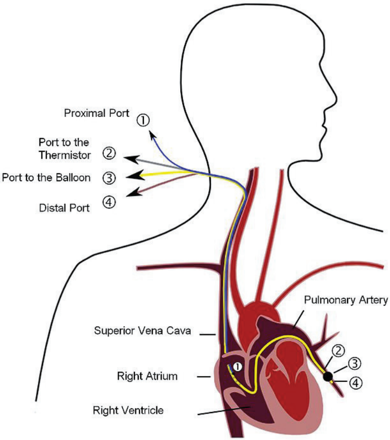


Figure 1: Pulmonary Artery Catheter (PAC), from Wikipedia

,

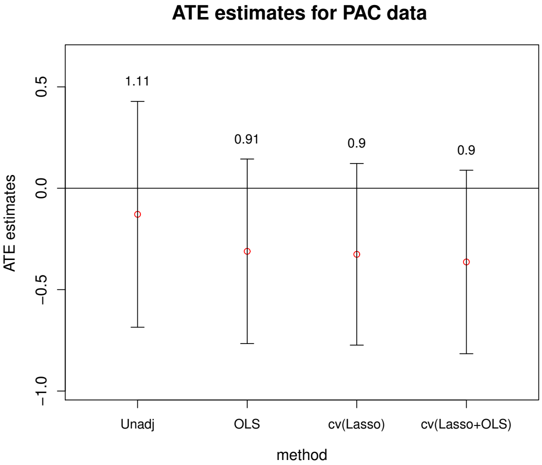
Motivation example: PAC-man

- Pulmonary Artery Catheter (PAC): monitoring device commonly inserted into critically ill patients
- Detecting complications, but invasive to patients and significant expenditure
- Question: does PAC have effect on patient survival?
- Observational study (Connors et al, 1996): PAC had an adverse effect on patient survival and led to increased cost of care

Motivation example: PAC-man data (Harvey et al, 2005)

- Randomized controlled trial on PAC, called PAC-Man: 65 UK intensive care units, 2001-2004
- 1013 patients: 506 treated and 507 control
- Outcome variable: quality-adjusted life years
- Observe many covariates including age, sex, some indicators, ...

PAC-man data: results



References

- Splawa-Neyman J, Dabrowska DM, Speed TP (1990). "On the Application of Probability Theory to Agricultural Experiments." Essay on Principles. Section 9. *Statistical Science*. 5(4): 465-472.
- Rubin DB (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies."
- Freedman DA (2008). "On regression adjustments to experimental data.". *Advances in Applied Mathematics* 40(2):180- 193.
- Freedman DA (2008). "On regression adjustments in experiments with several treatments.". *The Annals of Applied Statistics* 2(1):176-196.
- Lin W (2013). "Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique." *The Annals of Applied Statistics* 7:295-318.
- Tibshirani R (1994). "Regression Selection and Shrinkage via the Lasso." *Journal of the Royal Statistical Society B* 58:267-288.
- Belloni A, Chernozhukov V, Hansen C (2013). "Inference on Treatment Effects after Selection among High-Dimensional Controls". *The Review of Economic Studies* 81(2):608-650.

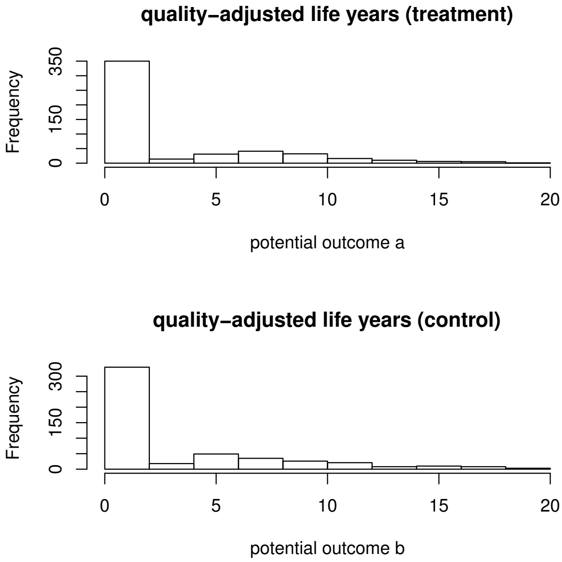
References

- Belloni A, Chernozhukov V, Fernsandez-Val I, Hansen C (2013). "Program evaluation with high-dimensional data." *arXiv preprint arXiv:1311.2645.*
- Tian L, Alizadeh A, Gentles A, Tibshirani R (2014). "A simple method for detecting interactions between a treatment and a large number of covariates." *Journal of the American Statistical Association* accepted.
- Rosenblum M, Liu H, En-Hsu Y (2014). "Optimal Tests of Treatment Effects for the Overall Population and Two Subpopulations in Randomized Trials, Using Sparse Linear Programming." *Journal of the American Statistical Association* 109(507):1216-1228.
- Connors AF et al. (1996). "The effectiveness of right heart catheterization in the initial care of critically Ill patients." *Jama* 276(11):889-897.
- Harvey S et al. (2005). "Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (PAC-Man): a randomised controlled trial." *Lancet* 366(9484):472-477.

References

- Chatterjee, A. and Lahiri, S. N. (2011). "Bootstrapping Lasso estimators". *J AM STAT ASSOC* 106:608-625.
- Chatterjee, A. and Lahiri, S. N. (2013). "Rates of convergence of the adaptive Lasso estimators to the oracle distribution and higher order refinements by the bootstrap". *ANN STAT* 41(3):1232-1259.
- Hanzhong, L. and Bin, Y. (2013). "Asymptotic properties of Lasso+mLS and Lasso+Ridge in Sparse High-dimensional Linear Regression". *ELECTRON J STAT* 7:3124-3169.
- Zhang, C. H. and Zhang, S. S. (2014). "Confidence interval for low-dimensional parameters in high-dimensional linear models". *J ROY STAT SOC B* 76(1):217-242.
- Van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2013). "On asymptotically optimal confidence regions and tests for high-dimensional models". *arXiv:1303.0518*
- Javanmard, A. and Montanari, A. (2013). "Confidence intervals and hypothesis testing for high-dimensional regression". *arXiv:1306.3171*

Histogram of potential outcomes



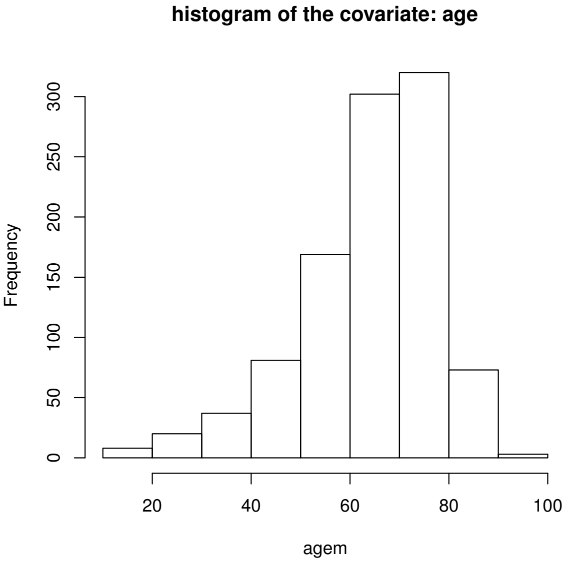
Selected covariates for PAC data

Table 2: Selected covariates for adjustment

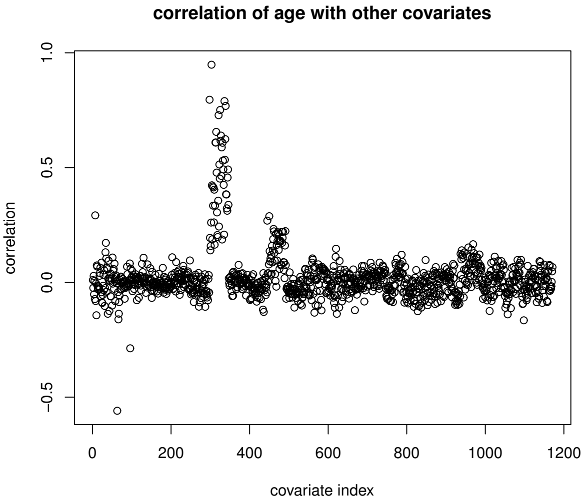
Method	T/C	Selected covariates
cv(Lasso+OLS)	T	age, p_death, age·age, age:p_death
cv(Lasso+OLS)	C	age, p_death, age·age, age:p_death, p_death:mech_vent
cv(Lasso)	T	pac_rate, age, p_death, age·age, p_death·p_death, region:im_score, region:systemnew, pac_rate:age, pac_rate:p_death, pac_rate:systemnew, im_score:interactnew, age:glasgow, age:systemnew, interactnew:systemnew, pac_rate:creatinine, age:mech_vent, age:respiratory, age:p_death, age:creatinine, interactnew:systemnew, interactnew:mech_vent, interactnew:male, systemnew:male, p_death:mech_vent, glasgow:organ_failure
cv(Lasso)	C	age, p_death, age·age, unitsize:p_death, pac_rate:systemnew, age:p_death, interactnew:mech_vent, p_death:mech_vent

T: treated; C: control. Covariate meanings: age (patient's age); p_death (baseline probability of death); mech_vent (mechanical ventilation at admission); region (geographic region); pac_rate (PAC rate in unit); creatinine, respiratory, glasgow, interactnew, organ_failure, systemnew, im_score (various physiological indicators).

Histogram of the covariate: age



Correlation



Proof Sketch

- Recall that

$$\widehat{ATE}_{\text{Lasso}} = \left[\bar{a}_A - (\bar{x}_A - \bar{x})^T \hat{\beta}^{(a)}\right] - \left[\bar{b}_B - (\bar{x}_B - \bar{x})^T \hat{\beta}^{(b)}\right]$$

- Since,

$$\begin{aligned}
&\sqrt{n}\left[\bar{a}_A - (\bar{x}_A - \bar{x})^T \hat{\beta}^{(a)}\right] \\
&= \sqrt{n}\left[\bar{a}_A - (\bar{x}_A - \bar{x})^T \beta^{(a)}\right] + \sqrt{n}(\bar{x}_A - \bar{x})^T(\beta^{(a)} - \hat{\beta}^{(a)})
\end{aligned}$$

- We have[2]

$$\begin{aligned}
&\sqrt{n}(\widehat{ATE}_{\text{Lasso}} - ATE) \\
= \quad & \sqrt{n}\left\{(\bar{a}_A - (\bar{x}_A - \bar{x})^T\beta^{(a)}) - (\bar{b}_B - (\bar{x}_B - \bar{x})^T\beta^{(b)}) - ATE\right\} \quad (1) \\
&+\sqrt{n}(\bar{x}_A - \bar{x})^T(\beta^{(a)} - \hat{\beta}^{(a)}) - \sqrt{n}(\bar{x}_B - \bar{x})^T(\beta^{(b)} - \hat{\beta}^{(b)}) \quad (2)
\end{aligned}$$

[2] Freedman DA (2008). On regression adjustments in experiments with several treatments. The Annals of Applied Statistics 2(1):176-196.

Proof Sketch

- Enough to show:

$$\sqrt{n}(\bar{x}_A - \bar{x})^T(\beta^{(a)} - \hat{\beta}^{(a)}) \to_p 0 \tag{3}$$

$$\sqrt{n}(\bar{x}_B - \bar{x})^T(\beta^{(b)} - \hat{\beta}^{(b)}) \to_p 0 \tag{4}$$

- By Hölder inequality,

$$|\sqrt{n}(\bar{x}_A - \bar{x})^T(\beta^{(a)} - \hat{\beta}^{(a)})| \leq ||\sqrt{n}(\bar{x}_A - \bar{x})||_\infty ||\beta^{(a)} - \hat{\beta}^{(a)}||_1$$

- Need to control

$$(I): \ ||\bar{x}_A - \bar{x}||_\infty \ \text{ and } \ (II): \ ||\beta^{(a)} - \hat{\beta}^{(a)}||_1$$

Bound (I): Cont

Lemma 2

Under the fourth moment condition on the covariates, if we let $c_n = \frac{(1+\tau)L^{1/4}}{p_A}\sqrt{\frac{2\log p}{n}}$, then as $n \to \infty$,

$$P\left(\|\bar{x}_A - \bar{x}\|_\infty > c_n\right) \to 0 \tag{5}$$

Thus, $||\bar{x}_A - \bar{x}||_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right)$.

Bound (II): Consistency of Lasso

- Proceed with the similar procedure with Lasso in linear regression, we can show

$$||\beta^{(a)} - \hat{\beta}^{(a)}||_1 = O_p\left(\frac{s\sqrt{\log p}}{\sqrt{n}} + s\delta_n\right) = o_p\left(\frac{1}{\sqrt{\log p}}\right)$$

- Both needs to show the following event happens w.h.p

$$\mathcal{L} := ||\frac{2}{n_A}\sum_{i\in A}(e_i - \bar{e}_A)(x_i - \bar{x}_A)^T||_\infty \leq \frac{\lambda_a}{2} \tag{6}$$

- What are the differences?

Bound (II): Consistency of Lasso

Difference is here

- Linear regression
  - A is fixed at $\{1, 2, ..., n\}$
  - e_i's are i.i.d zero mean Gaussian or Subgaussian random variable
  - Concentration inequality for i.i.d Gaussian or Subgaussian
- Neyman-Rubin model
  - e_i's are fixed number
  - A is a random set with fixed size
  - Concentration (Massart) inequality for sampling with replacement

Bound (I) + Bound (II)

- We have shown

$$||\beta^{(a)} - \hat{\beta}^{(a)}||_1 = O_p\left(\frac{s\sqrt{\log p}}{\sqrt{n}} + s\delta_n\right) = o_p\left(\frac{1}{\sqrt{\log p}}\right)$$

$$||\bar{x}_A - \bar{x}||_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right)$$

- Therefore,

$$\begin{aligned}||\sqrt{n}(\bar{x}_A - \bar{x})||_\infty ||\beta^{(a)} - \hat{\beta}^{(a)}||_1 &= \sqrt{n} O_p\left(\sqrt{\frac{\log p}{n}}\right) o_p\left(\frac{1}{\sqrt{\log p}}\right) \\ &= o_p(1)\end{aligned}$$